Addressing the Data Bottleneck in Implicit Discourse Relation Classification

Dissertation

zur Erlangung des akademischen Grades

eines Doktors der Philosophie

an den Philosophischen Fakultäten

der Universität des Saarlandes

vorgelegt von

Wei Shi

aus Hubei, China

Saarbrücken, 2020

Dekan der Fakultät P: Prof. Dr. Heinrich Schlange-Schöningen

Erstgutachter: Prof. Dr. Vera Demberg

Zweitgutachter: Prof. Dr. Josef van Genabith

Tag der letzten Prüfungsleistung: 19th November, 2020

Erklärung

Hiermit erkläre ich, dass ich die vorliegende Dissertation selbständig verfasst habe, und dass es meine eigene Forschung beschreibt. Keine anderen als die angegebenen Quellen und Hilfsmittel sind verwendet.

Declaration

I hereby declare that I composed this thesis entirely myself and that is describes my own research. I have not used any literature or materials other than the ones referred to in this thesis.

Wei Shi

Saarbrücken

Abstract

When humans comprehend language, their interpretation consists of more than just the sum of the content of the sentences. Additional logic and semantic links (known as coherence relations or discourse relations) are inferred between sentences / clauses in the text. The identification of discourse relations is beneficial for various NLP applications such as question-answering, summarization, machine translation, information extraction etc.

Discourse relations are categorized into implicit and explicit discourse relations depending on whether there is an explicit discourse marker between the arguments. In this thesis, we mainly focus on the implicit discourse relation classification, given that with the explicit markers acting as informative cues, the explicit relations are relatively easier to identify for machines.

The recent neural network based approaches in particular suffer from insufficient training (and test) data. As shown in Chapter 3 of this thesis, we start out by showing to what extent the limited data size is a problem in implicit discourse relation classification, and propose data augmentation methods with the help of cross-lingual data. And then we propose several approaches for better exploiting and encoding various types of existing data in the discourse relation classification task.

Most of the existing machine learning methods train on sections 2-21 of the PDTB and test on the section 23, which only includes a total of less than 800 implicit discourse relation instances. With the help of cross validation, we argue that the standard test section of the PDTB is too small to draw conclusions upon. With more test samples in the cross validation, we would come to very different conclusions about whether a feature is generally useful.

Second, we propose a simple approach to automatically extract samples of implicit discourse relations from multilingual parallel corpus via back-translation. After back-translating from target languages, it is easy for the discourse parser to identify those examples that are originally implicit but explicit in the back-translations. Having those additional data in the training set, the experiments show significant improvements on different settings.

Finally, having better encoding ability is also of crucial importance in terms of improving the classification performance. We propose different methods including a sequence-to-sequence neural network and a memory component to help have better representation of the arguments. We also show that having the correct next sentence is beneficial for the task within and across domains, with the help of the BERT (Devlin et al., 2019) model. When it comes to a new domain, it is beneficial to integrate external domain-specific knowledge. In Chapter 8, we show that with the entity-enhancement, the performance on BioDRB is improved significantly, comparing with other BERT based methods.

In sum, the studies reported in this dissertation contribute to address the data bottleneck problem in implicit discourse relation classification and propose corresponding approaches that achieve 54.82% and 69.57% on PDTB and BioDRB respectively.

Zusammenfassung

Wenn Menschen Sprache verstehen, besteht ihre Interpretation aus mehr als nur der Summe des Inhalts der Sätze. Zwischen Sätzen im Text werden zusätzliche logische und semantische Verknüpfungen (sogenannte Kohärenzrelationen oder Diskursrelationen) hergeleitet. Die Identifizierung von Diskursrelationen ist für verschiedene NLP-Anwendungen wie Frage-Antwort, Zusammenfassung, maschinelle Übersetzung, Informationsextraktion usw. von Vorteil.

Diskursrelationen werden in implizite und explizite Diskursrelationen unterteilt, je nachdem, ob es eine explizite Diskursrelationen zwischen den Argumenten gibt. In dieser Arbeit konzentrieren wir uns hauptsächlich auf die Klassifizierung der impliziten Diskursrelationen, da die expliziten Marker als hilfreiche Hinweise dienen und die expliziten Beziehungen für Maschinen relativ leicht zu identifizieren sind. Es wurden verschiedene Ansätze vorgeschlagen, die bei der impliziten Diskursrelationsklassifikation beeindruckende Ergebnisse erzielt haben. Die meisten von ihnen leiden jedoch darunter, dass die Daten für auf neuronalen Netzen basierende Methoden unzureichend sind. In dieser Arbeit gehen wir zunächst auf das Problem begrenzter Daten bei dieser Aufgabe ein und schlagen dann Methoden zur Datenanreicherung mit Hilfe von sprachübergreifenden Daten vor. Zuletzt schlagen wir mehrere Methoden vor, um die Argumente aus verschiedenen Aspekten besser kodieren zu können.

Die meisten der existierenden Methoden des maschinellen Lernens werden auf den Abschnitten 2-21 der PDTB trainiert und auf dem Abschnitt 23 getestet, der insgesamt nur weniger als 800 implizite Diskursrelationsinstanzen enthält. Mit Hilfe der Kreuzvalidierung argumentieren wir, dass der Standardtestausschnitt der PDTB zu klein ist um daraus Schlussfolgerungen zu ziehen. Mit mehr Teststichproben in der Kreuzvalidierung würden wir zu anderen Schlussfolgerungen darüber kommen, ob ein Merkmal für diese Aufgabe generell vorteilhaft ist oder nicht, insbesondere wenn wir einen relativ großen Labelsatz verwenden. Wenn wir nur unseren kleinen Standardtestsatz herausstellen, laufen wir Gefahr, falsche Schlüsse darüber zu ziehen, welche Merkmale hilfreich sind.

Zweitens schlagen wir einen einfachen Ansatz zur automatischen Extraktion von Samples

impliziter Diskursrelationen aus mehrsprachigen Parallelkorpora durch Rückübersetzung vor. Er ist durch den Explikationsprozess motiviert, wenn Menschen einen Text übersetzen. Nach der Rückübersetzung aus den Zielsprachen ist es für den Diskursparser leicht, diejenigen Beispiele zu identifizieren, die ursprünglich implizit, in den Rückübersetzungen aber explizit enthalten sind. Da diese zusätzlichen Daten im Trainingsset enthalten sind, zeigen die Experimente signifikante Verbesserungen in verschiedenen Situationen. Wir verwenden zunächst nur französisch-englische Paare und haben keine Kontrolle über die Qualität und konzentrieren uns meist auf die satzinternen Relationen. Um diese Fragen in Angriff zu nehmen, erweitern wir die Idee später mit mehr Vorverarbeitungsschritten und mehr Sprachpaaren. Mit den Mehrheitsentscheidungen aus verschiedenen Sprachpaaren sind die gemappten impliziten Labels zuverlässiger.

Schließlich ist auch eine bessere Kodierfähigkeit von entscheidender Bedeutung für die Verbesserung der Klassifizierungsleistung. Wir schlagen ein neues Modell vor, das aus einem Klassifikator und einem Sequenz-zu-Sequenz-Modell besteht. Neben der korrekten Vorhersage des Labels werden sie auch darauf trainiert, eine Repräsentation der Diskursrelationsargumente zu erzeugen, indem sie versuchen, die Argumente einschließlich eines geeigneten impliziten Konnektivs vorherzusagen. Die neuartige sekundäre Aufgabe zwingt die interne Repräsentation dazu, die Semantik der Relationsargumente vollständiger zu kodieren und eine feinkörnigere Klassifikation vorzunehmen. Um das allgemeine Wissen in Kontexten weiter zu erfassen, setzen wir auch ein Gedächtnisnetzwerk ein, um eine explizite Kontextrepräsentation von Trainingsbeispielen für Kontexte zu erhalten. Für jede Testinstanz erzeugen wir durch gewichtetes Lesen des Gedächtnisses einen Wissensvektor. Wir evaluieren das vorgeschlagene Modell unter verschiedenen Bedingungen und die Ergebnisse zeigen, dass das Modell mit dem Speichernetzwerk die Vorhersage von Diskursrelationen erleichtern kann, indem es Beispiele auswählt, die eine ähnliche semantische Repräsentation und Diskursrelationen aufweisen.

Auch wenn ein besseres Verständnis, eine Kodierung und semantische Interpretation für die Aufgabe der impliziten Diskursrelationsklassifikation unerlässlich und nützlich sind, so leistet sie doch nur einen Teil der Arbeit. Ein guter impliziter Diskursrelationsklassifikator sollte sich auch der bevorstehenden Ereignisse, Ursachen, Folgen usw. bewusst sein, um die

Diskurserwartung in die Satzdarstellungen zu kodieren. Mit Hilfe des kürzlich vorgeschlagenen BERT-Modells versuchen wir herauszufinden, ob es für die Aufgabe vorteilhaft ist, den richtigen nächsten Satz zu haben oder nicht. Die experimentellen Ergebnisse zeigen, dass das Entfernen der Aufgabe zur Vorhersage des nächsten Satzes die Leistung sowohl innerhalb der Domäne als auch domänenübergreifend stark beeinträchtigt.

Die begrenzte Fähigkeit von BioBERT, domänenspezifisches Wissen, d.h. Entitätsinformationen, Entitätsbeziehungen etc. zu erlernen, motiviert uns, externes Wissen in die vortrainierten Sprachmodelle zu integrieren. Wir schlagen eine unüberwachte Methode vor, bei der Information-Retrieval-System und Wissensgraphen-Techniken verwendet werden, mit der Annahme, dass, wenn zwei Instanzen ähnliche Entitäten in beiden relationalen Argumenten teilen, die Wahrscheinlichkeit groß ist, dass sie die gleiche oder eine ähnliche Diskursrelation haben. Der Ansatz erzielt vergleichbare Ergebnisse auf BioDRB, verglichen mit Baselinemodellen. Anschließend verwenden wir die extrahierten relevanten Entitäten zur Verbesserung des vortrainierten Modells K-BERT, um die Bedeutung der Argumente besser zu kodieren und das ursprüngliche BERT und BioBERT mit einer Genauigkeit von 6,5% bzw. 2% zu übertreffen.

Zusammenfassend trägt diese Dissertation dazu bei, das Problem des Datenengpasses bei der impliziten Diskursrelationsklassifikation anzugehen, und schlägt entsprechende Ansätze in verschiedenen Aspekten vor, u.a. die Darstellung des begrenzten Datenproblems und der Risiken bei der Schlussfolgerung daraus; die Erfassung automatisch annotierter Daten durch den Explikationsprozess während der manuellen Übersetzung zwischen Englisch und anderen Sprachen; eine bessere Repräsentation von Diskursrelationsargumenten; Entity-Enhancement mit einer unüberwachten Methode und einem vortrainierten Sprachmodell.

Acknowledgments

Choosing to pursue a Ph.D. here in Saarland Univeristy is one of the best and wisest decisions I have ever made. Looking back over the last more than four years, it is such a long climb and a tough journey that eventually approaches its ending. And there is a quite long list of people whom I would like to appreciate.

First of all, my sincere gratitude goes to my supervisor, Prof. Dr. Vera Demberg, for being a fantastic and insightful advisor. Thank you in the first place for inviting me to this country and continue my academic career, when I was a little bit confused about the future. You have lead and gathered so many awesome colleagues and created such nice research environments. Every research discussion and planning with you have benefited me greatly. Your rigorous and diligent show us what a qualified researcher should be like. Most importantly, thank you for listening to my ideas, being flexible about my agenda and providing adequate guidance.

I am grateful to the Collaborative Research Center SFB 1102 on "Information Density and Linguistic Encoding" (under German Research Foundation) who funded this research. My thanks also go to all the members of SFB 1102, especially the coordinators Marie-Ann Küne and Patricia Borrull Enguix, for all your assistance. It is my pleasure to work with all the smart people I met in Saarbrücken. I would like to thank all the (ex-)colleagues: Alessandra Zaarcone, Anupama Chingacham, David Howcroft, Ernie Chuanyi Zhang, Frances Yung, Gabriele Reibold, Jorrig Vogels, Katja Häuser, Katja Kravtchenko, Margarita Ryzhova, Marjolein van Os, Merel Scholman, Pratik Bhandari, Tony Xudong Hong, William Blaco and many others. I am grateful to have your company along the path and have enjoyed all the board game nights and discussions we had together.

Acknowledgements also go to the friends I met during my PhD, Dawei Zhu, Fangzhou Zhai, Yanzhe Guo and Yue Fan. Thank you for all the barbecue parties on the balcony, badminton practices, all the wins (and failures) in the Age of Empires game and research discussions upon computers, languages, visions and etc.. You made the journey less an agony and full of memorable experiences and happiness.

My deep thanks for my family: my parents and sisters, for being supportive whenever I needed help. Your everlasting supports helped me achieve one and another goals.

Finally, but foremost, I would like to thank my wife, Hao Wu. I would not be able to finish my studies without you. Thank you for always being on my side. Words are powerless to express my gratitude to you. What's past is prologue, I look forward to new experiences and adventures with you in the future.

Contents

Chapter 1

Introduction

The comprehension of language by humans are not simply the combination of interpretation of isolated and unrelated sentences or clauses, instead humans assign meaning to the sentences by adding logic links between the clauses. In this way, a piece of text is often being understood by connecting to other text units from its context. These units can be surrounding clauses, sentences or even paragraphs. These logic and semantic links between clauses are also known as *coherence / discourse relations*. Discourse relations between clauses also affect and add new interpretation of the text in many ways, let's consider the following example:

1. *Countries implement necessary quarantines and social distancing practices after the global spread of the coronavirus. The world economy is expected to have a large recession in 2020.*

In the first sentence, two events have been connected by the temporal connective *after*, which indicates that countries implement actions *after* the spread of the virus. But it can also be interpreted that the measures implemented by countries are *caused* by the global spread of the virus. What's more, without any explicit cues like the word *after*, the cause relation between the spread of virus and the economic recession can also be easily inferred.

If readers are not able to construct the relationships between sentences and simply sum the meanings of sentences solely, they will fail to fully understand the context of the text. Hence, discourse relation is crucial to natural language understanding in general.

In this chapter, we first give a general introduction to the basic concepts of computational discourse relation parsing. And then we briefly introduce the research questions discussed in this dissertation which is followed by the contributions to those questions. At last, we have an overview of the structure of this dissertation.

1.1 Computational discourse relation parsing

Even when a text is well-structured, finding the discourse relationships that hold texts together automatically is difficult. The process of discourse-level analysis may lead to a number of natural language process tasks: connective identification, discourse segmentation, discourse relation classification.

The Penn Discourse Treebank (Prasad et al., 2008) adopts a binary predicate-argument view on discourse relations, where the connective acts as a predicate that takes two text spans as its arguments. The span to which the connective is syntactically attached is called *Argument 2*, while the other one is called *Argument 1*. Thus the first step to analyze discourse is to identify the connectives.

Discourse relations are held between two attributes. Given a paragraph of raw text, identifying the spans for attributes is called discourse segmentation. As for explicits, having the connective attached to Arg2 makes the task relatively easier because only Arg1 needs to be extracted. However, for implicits, both arguments spans need to be identified.

After having both arguments, how to encode and classify the discourse relation between them leads us to the task of discourse relation classification. In this dissertation, we focus on the implicit discourse relation classification where the discourse relations are not signaled explicitly by discourse connectives. Previous work has shown that implicit discourse relation classification has been the bottleneck of discourse parsing due to the difficulty in better representing the semantic and syntactic information of the arguments.

1.2 Research questions

With the booming successes of deep learning methods in natural language processing in recent years, lots of neural network models have been proposed for implicit discourse relation classification. However, due to the difficulties of manually annotating discourse relation data, most of the models are trained and evaluated on the single largest available discourse relation corpus available for now, the Penn Discourse Treebank, with some conventional settings. In this dissertation, we try to answer the following questions:

1. Is the currently most-used dataset large enough for the machine learning methods to train on? Is it risky to draw conclusions about whether the inclusion of certain features constitute a genuine improvement depending on the results on a small test set?

2. Manually annotating discourse relations, implicit ones in particular, are very expensive and time-consuming. Is there a way to acquire automatically annotated implicit discourse relation data with high confidence?

3. With limited number of data, learning the surface cues is obviously not adequate. How to have better understanding of how arguments relate to one another and to have better semantic representations are of crucial importance for implicit discourse relation classification.

4. Having good encoding only does part of the job, a good implicit discourse relation classifier should be competent in being able to encode discourse expectation and learn typical temporal event sequences, causes, consequences etc. for all kinds of events. This motivates us to figure out whether the next sentence prediction subtask in BERT (Devlin et al., 2019) is really helpful or not in capturing the upcoming events.

5. For neural network models, the domains matter a lot. The differences in vocabulary and writing style across domains can cause state-of-the-art supervised models to dramatically increase in error. The gap between different domains, like Wall Street Journal (economic journals) and PubMed (biomedical journals) dataset, has great influences on the models. How to shift and reduce the impacts of domain discrepancies with the pre-trained language models and external entity knowledge?

These are important questions for the task of implicit discourse relation classification, especially for the neural network models. In this dissertation, we tackle these questions in three pathways: data argumentation, representation modeling and entity enhancement. Specifically, we first expose the risky idea of drawing conclusions with the results from a limited number of test samples and advocate to use cross validation instead. Then we propose to use back-translation methods on multi-lingual data to expand the scale of annotated training data, which later also motivates us to use sequence to sequence model to better represent the relational arguments. Last but not the least, we try to use unsupervised methods and also entity enhancement techniques with pre-trained language models.

1.3 Dissertation contributions

This dissertation is concerned with using machine learning methods to assign the discourse relation to a pair of sentences with no explicit discourse connective in between. We tackle this task from different angles inspired by the weaknesses of previous proposed methods, and propose new approaches respectively. The major contributions of this dissertation are summarized as follows.

1.3.1 Limited data size problem

In recent studies, various classes of features are explored to capture lexical and semantic regularities for identifying the sense of implicit discourse relations, including linguistically informed features like polarity tags, Levin verb classes, length of verb phrases, language model based features and constituent features etc.. Most of them are trained and tested on the PDTB, in which there are only a dozen instances for some of the second-level relations. The PDTB is split from the Penn Treebank (Marcus et al., 1993), which has a lot more instances to learn from for the parsing community. Conclusions about the effectiveness of including certain features are made depending on the performances on the conventional most-used test set, which has only less than 800 implicit relations. In this work, we aim to demonstrate the degree to which conclusions would depend on whether one evaluates on the standard test section only, or performs cross validation on the whole dataset for the

second-level discourse relation classification.

We employ simple Long Short-Term Memory networks that concatenate surface features to predict the implicit discourse senses, given both the relational arguments. Our experiment results suggest that it comes to very different conclusions if actually running the cross-validation experiments, which means that the standard test section of the PDTB is way too small to draw conclusions about whether a feature is generally beneficial to this task or not, especially when we use a relatively larger label sets. We run a large risk of drawing incorrect conclusions about which features are helpful if we only stick out our small standard test set. In this work we argue in favor of significance testing with cross validation, as opposed to boot strapping methods that only use the standard small test set. This is the first work that systematically evaluates the effect of the train/test split for the implicit discourse relation classification task on PDTB.

1.3.2 Data augmentation with multi-lingual back-translation

With the increasing number of parameters to be trained, most of the proposed neural network models for implicit discourse relation classification suffer from the shortage of labeled data. While manually annotating implicit discourse relations requires professional linguistic knowledge, is time consuming and also expensive, we hereby address the problem by procuring additional training data from parallel corpora. When human translate a text, they sometimes add connectives (also know as explicitation process). We automatically back-translate it into an English connective and use it to offer an explicit label to the original implicit English with high confidence.

The pipeline works as follows. Firstly we back-translate the target French sentence from corpora that are mostly used in machine translation task into English using a pre-trained machine translation system. Then with the help of an end-to-end discourse parser, we parse both the original and back-translated English sentences. The parser will output a list of explicit relations including the discourse relation tags and argument spans. Since all the implicit instances are consecutive sentences in the PDTB scheme, we follow this rule and then identify the implicit-to-explicit discourse relation alignments according the outputs of end-to-end parser. We extract those sentence pairs that hold implicit relation in the original

English but explicit in the back-translations. In the end we label the source English sentence pairs with the relation tag of the explicit relation in the back-translated target text.

However, this method still suffers from the fact that typical sentence-aligned corpora may have some sentences removed and make the consecutive sentences no longer coherent to get inter-sentential discourse instances. In addition, using a single language pair (English-French in this case) might not be sufficient to get instances with high confidence. In order to solve these questions, we expand the pipeline described above with more preprocessing steps and more language pairs. With the majority votes from different language pairs, the mapped implicit labels are more reliable.

1.3.3 Better representation of relation arguments

Given that there is no connectives acting as informative cues for implicit discourse relation classification, the difficulties of this task has shifted into how to effectively encode the relational arguments. We here propose a new model, which consists of a classifier and a sequence-to-sequence model which is trained to generate a representation of the discourse relation arguments by trying to predict the arguments including a suitable implicit connective. The whole method is trainable because such implicit connectives have been annotated as part of the PDTB corpus. This novel secondary task forces the internal representation to more completely encode the semantics of the relation arguments and to make a more fine-grained classification. To further capture common knowledge in contexts, we also employ a memory network to get explicit context representation of contexts training examples. For each test instance, we generate a knowledge vector by weighted memory reading. Experimental results show that with the context memory, the model can facilitate the discourse relation prediction by choosing examples that share similar semantic representation and discourse relation.

The successful use of memory network means that having the relevant context is beneficial for implicit discourse relation classification. However, it can only provide some background knowledge and give hints as to what topic the instance possibly is about and what coherence relation may be present. It is clear that models cannot learn all these diverse relations from the limited amounts of available training data. A more general representation of discourse

expectations should also be vital and learnt.

After seeing all those successes in various NLP tasks made by the bidirectional encoder representation from transformers (BERT) proposed by Devlin et al. (2019), we see that the next sentence prediction ,which has been used in BERT as a sub-task, is a very good fit for the implicit discourse relation classification. It allows the model to be aware of what typical causes, consequences, events or contrasts are coming up in the next sentence. To have better understanding about what role the next sentence prediction task plays in model, we try the BERT model with and without the sub-task within and across domains. In addition, in-domain continue pre-training with BERT has been proven very useful in improving the performance of this task. With the continue training on the new domain data, the model shows very competitive ability in shifting across domains, compared with the whole-in-domain training model like BioBERT (Lee et al., 2020).

1.3.4 Entity-enhanced pre-trained language model

In the last work, we show that with a small amount of in-domain data for continue pre-training, the BERT model demonstrates very competitive ability in shifting across domains. However, comparing with the BioBERT which is pre-trained with gigantic in-domain data, the improvement brought by the in-domain raw texts pre-training is very limited. One of the reasons is that for a new domain such as biomedical, the entities may have different formats or appear very rarely. This sparsity problem makes it very hard for neural network language models to encode.

In this work, we first use the information retrieval system to extract the SPO (Subject, Predicate, Object) triples from the explicit discourse instances. We assume that if both discourse instances are talking about the same entities, there is high possibility that they share some similarities in the sense of discourse relations. The experiments show that even with the unsupervised majority voting system, the proposed method achieves comparable results comparing with the BERT models that are trained across domains. In addition, we employ the recent proposed approach K-BERT which injects the domain-specific entity knowledge into the pre-trained model. With the extra SPO triples we extract with the knowledge graph system, the classification performance outperforms the BioBERT with 2% accuracy and be-

comes the new state of the art result on the BioDRB.

1.4 Overview of the dissertation

This dissertation is organised into nine chapters.

- **Chapter 2** provides a general introduction to discourse relation including some basic concepts of discourse relation and as well as different theories in the recent decades. Furthermore, we look into the available resources for computational study of discourse relations and give an overview of the Penn Discourse Treebank, which is a discourse-level annotation atop the Penn Treebank. And then we review the recent proposed neural modeling techniques and talk about the recent machine learning attempts and advances in the task of discourse modelling.

- In **Chapter 3**, we talk about the problem of limited number of training data for the task. We argue that the standard test section of the PDTB is too small to draw conclusions about whether a feature is generally useful or not, especially when using a larger label set, and we run large risk of having incorrect conclusions if we stick to the small, most-used standard test set. Instead, we advocate to make full use of the whole dataset by using cross validation.

- In **Chapter 4**, we propose a new pipeline to automatically annotating original English sentences with the help of back-translated connectives of French sentences. After back-translating the French sentences from the English-French parallel corpus, we identify the explicit connectives in the back-translation and label the original English with high confidence. In this way, we expand the training data to a significant size.

- In **Chapter 5**, we expand the idea from Chapter 4 to multiple language pairs, which means that each implicit discourse relation instance from the original English is annotated with multiple labels. With the majority vote from those labels, instances with ambiguous relations and as well as the disagreements between different human translators have been filtered out. In addition, we use paragraph-aligned sentences to keep

the topic of both arguments consistent and also use statistical machine translation system to get stabler translation of connectives.

- In **Chapter 6**, we propose a novel sequence-to-sequence model to encode the relational arguments for implicit discourse relation classification. It consists of a classifier, a memory component and a seq2seq model which is trained to generate a representation of the arguments by trying to predict the relational arguments including a suitable implicit connective. The additional secondary task of explicitation forces the internal representation to more completely encode the semantics of the relational arguments and has been proven beneficial by the experimental results on different test settings.

- **Chapter 7** discusses the necessity of capturing what events are expected to cause or follow each other in extracting better representations of the relational arguments. We look into what role the next sentence prediction sub-task plays in the recently proposed model in Devlin et al. (2019). We show that BERT has very good ability in encoding the semantic relationship between sentences with its "next sentence prediction" task in pre-training. In particular, with several epochs of continuous pre-training on the in-domain data, it also shows very good capability in domain shifting and outperforms the recent state of the art system with a substantial margin across domain.

- In **Chapter 8**, we propose a pipeline with the information retrieval and knowledge graph system to extract the most relevant SPO triples, given an implicit discourse instance as a query. With different matching strategies, we first use an unsupervised method for the relation prediction and achieve competitive results on BioDRB. We then employ the recent K-BERT method along with our extracted SPO triples as external entity knowledge, and achieve the state of the art results, outperform the BioBERT with a significant margin.

- Lastly, **Chapter 9** summarizes the work in this dissertation and outlines a number of future directions for the task of implicit discourse relation classification.

1.5 Relevant publications

- Shi, W., & Demberg, V. (2017). On the need of cross validation for discourse relation classification. In Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers (pp. 150-156). Valencia, Spain

- Shi, W., Yung, F., Rubino, R., & Demberg, V. (2017). Using explicit discourse connectives in translation for implicit discourse relation classification. In Proceedings of the Eighth International Joint Conference on Natural Language Processing (Volume 1: Long Papers) (pp. 484-495). Taipei, Taiwan

- Shi, W., & Demberg, V. (2019). Learning to Explicitate Connectives with Seq2Seq Network for Implicit Discourse Relation Classification. In Proceedings of the 13th International Conference on Computational Semantics-Long Papers (pp. 188-199). Gutenberg, Sweden.

- Shi, W., Yung, F., & Demberg, V. (2019). Acquiring Annotated Data with Cross-lingual Explicitation for Implicit Discourse Relation Classification. In Proceedings of the Workshop on Discourse Relation Parsing and Treebanking 2019 (pp. 12-21). Minneapolis, MN, USA

- Shi, W., & Demberg, V. (2019). Next Sentence Prediction helps Implicit Discourse Relation Classification within and across Domains. In Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP) (pp. 5794-5800). Hongkong, China

Chapter 2

Background

This chapter briefly presents an overview of some basic concepts of discourse relation and previous work that is relevant to the work reported in this thesis. I start with an introduction to the definition of implicit discourse relation and how they are categorized and annotated in the Penn Discourse Treebank (PDTB), followed by some recently proposed neural network models in general natural language process, that are essential as some techniques have partially been used in this thesis. Finally I survey previous approaches in this task, including traditional feature-based methods and recent successes of neural network models.

2.1 Introduction to discourse relation

When human comprehend language, their interpretation consists more than just the sum of the contents of the sentences. Additional semantic relations (known as coherence relations or discourse relations) are inferred between sentences in the text.

To better understand the notion of coherence, consider the following example from Tannen et al. (2015). There were two signs in a swimming pool. One of them said, *Please use the toilets, not the pool*, the other sign said, *Pool for members only.* Each sign is reasonable enough if we read them separately. But when the two sentences are read as if they were part of a single discourse, the second sentence forces a reinterpretation of the first one. With the relationship

between sentences, new meanings are created.

Discourse relations describe the logical relation between two sentences/clauses, they reveal the structural organization of text and allow for additional inferences. The identification of discourse relations is beneficial for various downstream NLP applications such as question-answering (Liakata et al., 2013; Jansen et al., 2014), summarization (Maskey and Hirschberg, 2005; Yoshida et al., 2014; Gerani et al., 2014), machine translation (Guzmán et al., 2014; Meyer et al., 2015) and information extraction (Cimiano et al., 2005). In recent years, the task of discourser relation parsing has drawn increasing attention, including two CoNLL shared tasks. (Xue et al., 2015, 2016).

Discourse relations in texts are sometimes expressed with a connective (e.g., *but, because, however*, that are referred to as explicit discourse relation. However, connectives are often absent, while a discourse relation is still inferred. These are called implicit relations. As shown in the following example.

1. *[The federal government suspended sales of U.S. savings bonds.]*$_{Arg1}$ ***because*** *[Congress hasn't lifted the ceiling on government debt.]*$_{Arg2}$

   — Explicit, Contingency.Cause.Reason - wsj_0008

2. *["I believe in the law of averages," declared San Francisco batting coach Dusty Baker after game two.]*$_{Arg1}$ *(**Implicit = accordingly**) ["I'd rather see a so-so hitter who's hot come up for the other side than a good hitter who's cold."]*$_{Arg2}$

   — Implicit, Contingency.Cause.Result - wsj_2202

These two examples are annotated by the Penn Discourse Treebank, we will talk about PDTB in detail in the next section. The first example shows an example of explicit relations with a connective originally present in the text. With the explicit discourse connective "because", it is easy to know that the discourse relation is Contingency.Cause. However, if the connective is absent, as is the case in the second example, identifying the relation is more difficult. One way to identify the relation is to insert a connective, in this case the annotators inferred the connective "accordingly" that most intuitively connects Arg_1 and Arg_2. It is relatively easy for human to correctly infer the relation with the semantic and syntactic information

of the arguments, but not for machines. How to encode and get good representation of the arguments becomes the key point of this task, which also makes the implicit discourse relation classification very challenging and represents a bottleneck of the entire discourse parsing system.

According to the work of linguists over decades, a few theories of discourse relation have been developed, including the theory of coherence and coreference by Hobbs (1979), the rhetorical structure theory (RST) by Mann and Thompson (1988), the discourse structure theory by Grosz and Sidner (1986), the segmented discourse representation theory (SDRT) by Asher et al. (2003). Unfortunately, not so much consensus has been obtained to date in terms of the number and type of relation senses that should be considered in a standard discourse analysis (Asr, 2015). Among the few theories along with annotated corpora that illustrate the theories and allow for the computational aspect of discourse analysis, the RST-DT (Carlson et al., 2001) and PDTB (Prasad et al., 2008) are the most famous and frequently used corpora.

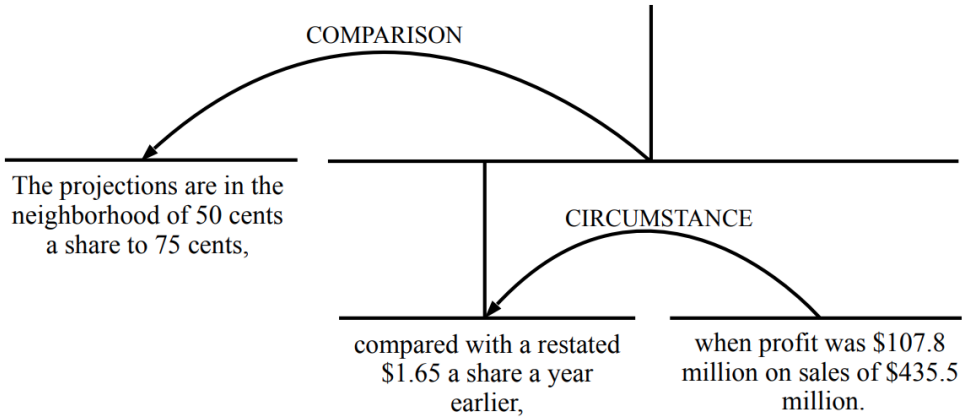

Figure 2.1: An example of RST discourse structure. (Taken from Ji and Eisenstein (2014)

Rhetorical Structure Theory Discourse Treebank (RST-DT)

The framework that is used to annotate RST-DT (Carlson et al., 2001) is based on the Rhetorical Structure Theory proposed by Mann and Thompson (1988). It adopts a tree as the structure that underlines relationships within the units of text. In terms of tree structure, the leaves of the tree correspond to text fragments that represent the minimal units of the discourse is called *elementary discourse units* (EDUs). The internal nodes of the tree correspond to contiguous text spans. And each node is characterized by its *nuclearity*: a *nucleus* indicates

Attribution	attribution, attribution-negative
Background	background, circumstance
Cause	cause, result, consequence
Comparison	comparison, preference, analogy, proportion
Condition	condition, hypothetical, contingency, otherwise
Contrast	contrast, concession, antithesis
Elaboration	elaboration-additional, elaboration-general-specific, elaboration-part-whole, elaboration-process-step, elaboration-object-attribute, elaboration-set-number, example, definition
Enablement	purpose, enablement
Evaluation	evaluation, interpretation, conclusion, comment
Explanation	evidence, explanation-argumentative, reason
Joint	list, disjunction
Manner-Means	manner, means
Topic-comment	problem-solution, question-answer, statement-response, topic-comment, comment-topic, rhetorical-question
Summary	summary, restatement
Temporal	temporal-before, temporal-after, temporal-same-time, sequence, inverted-sequence
Topic change	topic-shift, topic-drift
Structural	textual-organization, span, same-unit

Table 2.1: Tagset of discourse relations in RST-DT (Carlson and Marcu, 2001).

a more essential unit of information, while a *satellite* indicates a supporting or background unit of information. Each node is characterized by a rhetorical relation that holds between two or more non-overlapping, adjacent text spans, as shown in Figure 2.1. The arrow originates at the *nucleus* and points to the satellite with a discourse relation on the top. Relations in RST-DT can be semantic, intentional, or textual nature, and are explicitly classified into these three categories. Carlson et al. (2001) distinguish 78 relation labels, partitioned into 16 classes and share some type of rhetorical meaning, more details please refer to Table 2.1 and the annotation manual (Carlson and Marcu, 2001).

However, RST-DT discourse segments or relational arguments can vary between a phrase and a paragraph and the discourse connectives are only used for determining the boundaries of the discourse segments, i.e. they do not have an official status in determining the type of discourse relation (Asr, 2015).

With about 50k relations instances, PDTB is the largest resource of discourse relation in size. And the text comes from Wall Street Journal that have been also manually annotated for syntax in the Penn Treebank project (Marcus et al., 1994). It also has few labels compared with RST-DT introduced above. What's more, PDTB annotators have also achieved a relatively good level of inter-annotator agreement. Given all these factors, recent work

that focus on computational modelling of discourse relation prefer using another annotated corpora PDTB, we will give an overview of PDTB next.

2.2 Overview of the Penn Discourse Treebank

The PDTB 2.0 corpus (Prasad et al., 2008) is the largest manually annotated discourse relation corpus available at the moment. The framework that was used to annotate the corpus is referred to as PDTB as well and has been used to create new corpora in other languages such as Arabic (Al-Saif and Markert, 2010), Italian(Tonelli et al., 2010) and Chinese (Zhou and Xue, 2015). It covers the set of one million word *Wall Street Journal* (WSJ) articles in the Penn Treebank (PTB) (Marcus et al., 1994), which is much larger than the previous existing RST-DT corpus. Unlike RST-DT, PDTB does not have the smallest textual units. It adopts a binary predicate-argument view on the discourse relations, where the connective acts as a predicate that takes two text spans as its arguments.

PDTB distinguishes between explicit and implicit discourse relations depending on whether it is marked with a discourse connective. When annotating implicit relations, annotators are asked to insert a connective they think would best fit and annotate the coherence relation with the inserted connective. Relations in the PDTB have two and only two arguments, referred as *Arg1* and *Arg2*. These arguments can be continuous or discontinuous. In the case of explicit relations, the argument that is syntactically bound to the connective is labeled as *Arg2*, while the other one is *Arg1* which may be adjacent or non-adjacent with *Arg2*. However, implicit discourse relations have only been annotated between adjacent sentences with paragraphs, as well as between complete clauses delimited by a semi-colon (";") or colon (":").

As illustrated in Figure 2.2, 43 discourse relations have been distinguished in PDTB 2.0.The labels are organised in a hierarchy consisting of three levels:

(i) *class* is the top level, which contains the four major semantic classes: TEMPORAL, CONTINGENCY, COMPARISON and EXPANSION. Temporal is used when the events or situations in Arg1 and Arg2 are related temporally. The relation belongs to contingency when one argument causally influences or causes the other. When the events in Arg1 and Arg2 are

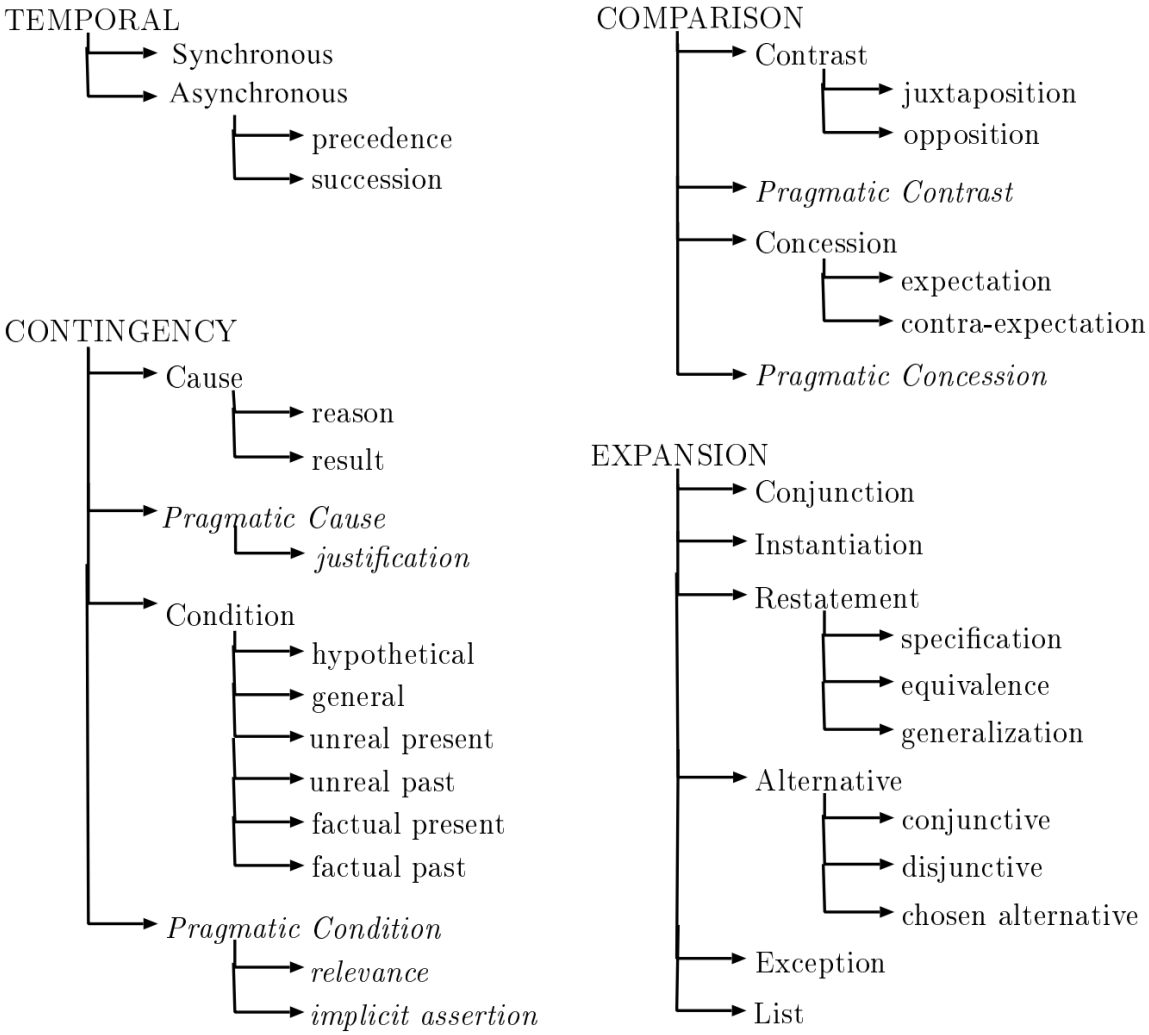

Figure 2.2: The Hierarchy of sense tags in PDTB (Prasad et al., 2008).

compared to highlight the difference, it is labeled as a Comparison and otherwise it is called Expansion if one argument is to expand the other one semantically or in discourse.

(ii) *type* is the second level which refines the semantic of the *class* level. For example, there are four *types* defined under the Comparison *class*: Contrast, Pragmatic Contrast, Concession and Pragmatic Concession.

(iii) *sub-type* is the most fine-grained level, which defines the semantic contribution and also direction of each argument, e.g. reason and result in Contingency.Cause.

If the annotator was uncertain of the more fine-grained senses of subtype, s/he could choose the higher level, which means that not every discourse relation instance in PDTB has all the three level relations. Conventionally, existing work on discourse parsing tend to evaluate their models either on the first-level 4-way classification (Pitler et al., 2009; Rutherford and Xue, 2014; Chen et al., 2016; Shi and Demberg, 2019a) or the second-level 11-way classification (Lin et al., 2009; Ji and Eisenstein, 2015; Qin et al., 2016a, 2017; Shi and Demberg, 2019b).

However there are three additional labels where an implicit discourse connective could not be inserted. A connective cannot always be inserted when there is no discernible coherent relation between two sentences (NoRel). This label is assigned when adjacent sentences don't stand in a direct relation to on another because they belong to two different discourse segments. When the sentences do coherent together but via only cohesion, like talking about the same entity instead of a definable coherent relation, they have been annotated as EntRel, which stands for *Entity Relation*. AltLex (*Alternative Lexicalization*) applies to examples that there is already another explicit expression presented and the insertion of a connective leads to a perception of relation redundancy.

In total, there are 40600 instances annotated from 25 sections in PDTB 2.0, of which there are 18459 explicit, 16053 implicit and 6088 others (AltLex, EntRel and NoRel). Given that PDTB is now the largest available corpus in discourse relation parsing and each instance have two arguments, which is easier for modeling, more and more existing work have been using it as the gold standard label for the evaluation of the proposed methods. In this thesis, we also mainly use the conventional PDTB for evaluation to make our results comparable to the existing work.

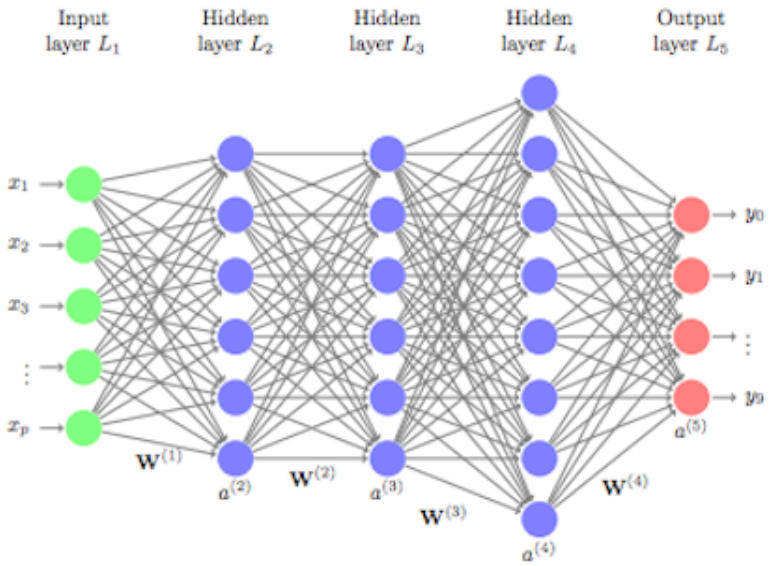

Figure 2.3: Mutilayer fully connected feedforward neural network.
Source: `http://uc-r.github.io/feedforward_DNN`

2.3 Neural modeling in natural language process

The recent decade has witnessed remarkable success by the neural networks in a lot of areas. It has become the industry-standard algorithm that achieves as low as 9% word error rate in speech recognition, boosts the machine translation performances and has been broadly used in many computer vision tasks as well.

However, compared with computer vision and speech recognition, there are relatively larger spaces to achieve good results on many tasks in natural language process. In this section, I will briefly introduce the recently proposed and mostly used neural network models by researchers in NLP.

FeedForward neural network

Deep feedforward network, also known as multilayer perceptrons, are the foundation of most deep learning models. The decision flow, as occurs in the single neuron, is unidirectional, advancing from the input to the output in successive layers, without cycles or loops. In a vanilla feedforward neural network, inputs are first transformed to a feature vector specifically designed for the task. Two operations are alternatively applied to the vectors: linear transformation and non-linear activation functions (sigmoid, tangent etc.), which results in a representation of the inputs and can be used for classification or the input

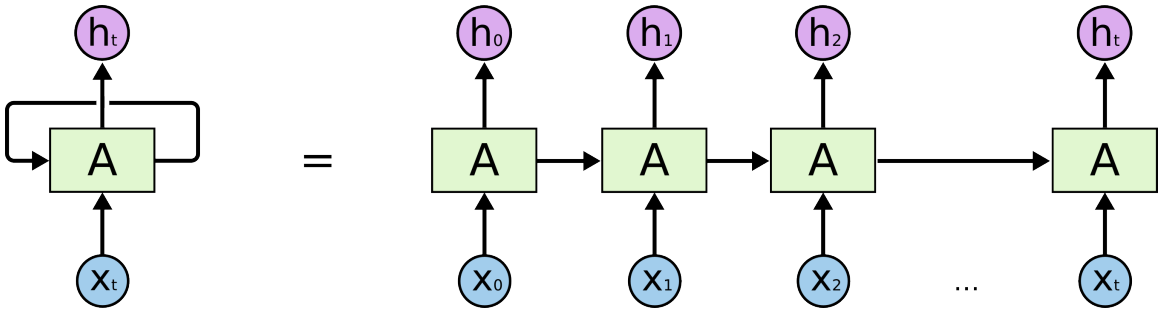

Figure 2.4: Recurrent Neural Networks.
Source: https://colah.github.io/posts/2015-08-Understanding-LSTMs/

of another round of linear and non-linear transformation, as shown in Figure 2.3. Formally, given the input $X \in R^d$, we derive the representation H_i with single layer feedforward units follow the below equation:

$$H_i = \sigma(W_i \cdot X + b_i) \tag{2.1}$$

, where the W_i denotes the weight matrix of linear transformation with the size of $k * d$ and b_i is a bias vector. The non-linear activation function σ is usually chosen to be tanh or sigmoid. For multilayer feedforward network, the X should be replaced with H_{i-1}, the output vector from previous layer.

Recurrent neural network (RNN)

It is easy to notice that the feedforward network can only take inputs with the fixed size, given that the size of linear transformation matrix is predetermined. However, for a time sensitive inputs like natural language or speech, it is vital to have a model that can be flexible in dealing with the time series. A recurrent neural network is designed to take a series of input with no predetermined limit on the size. What's more important, it remembers the past and it's decisions are influenced by what it has learnt from the past. As illustrated in Figure 2.4, there is a loop which allows information to be passed from one step of the network to the next. In other words, the hidden state of the current time step h_t is a function of the current time step input X_i and the hidden state of the previous time step h_{t-1}. It can be written in equation as follows:

$$h_t = \sigma(W \cdot h_{t-1} + U \cdot X_i + b) \tag{2.2}$$

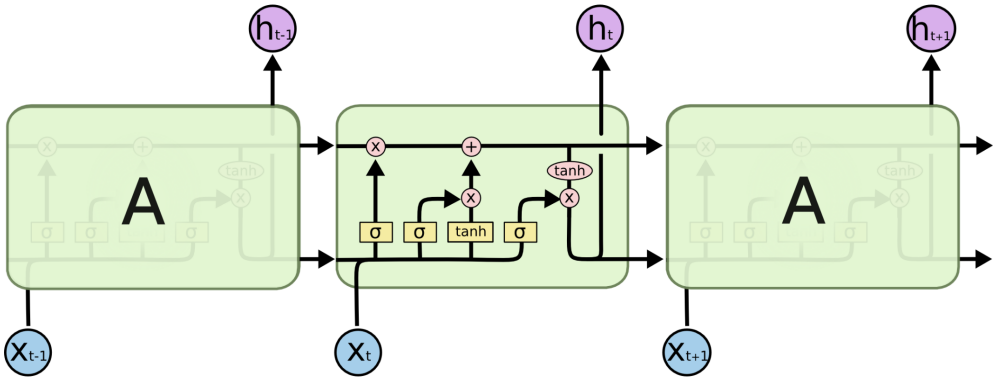

Figure 2.5: The repeating module of Long short-term memory cell.
Source: `https://colah.github.io/posts/2015-08-Understanding-LSTMs/`

In each time step, all the weight matrix W and U are shared, which makes the framework recurrent.

In order to minimize the cost function, neural network models are usually optimized with a gradient-based method as there is no closed-form solution. As for RNN, back-propagation through time (BPTT) is involved, in which the gradient calculation involves the whole sequence of hidden states and outputs. It is an efficient way for training without large scale of harms to the performance. Unfortunately, it has been observed by Bengio et al. (1994) that it is difficult to train RNNs to capture long-term dependencies because the gradients tend to either vanish (most of the time) or explode (rarely, but with severe effects), due to the small (gradient vanishing) or large (gradient explosion) values in the matrix and multiple matrix multiplications. This makes gradient-based optimization methods struggle (Chung et al., 2014).To address this problem, two RNN-based variants, Long short-term memory (LSTM) network (Hochreiter and Schmidhuber, 1997) (Figure 2.5) and Gated Recurrent Units (GRU) (Cho et al., 2014) are introduced. They are designed with more sophisticated gate layers rather than an usual activation function, consisting of affine transformation (linear transformation and translation), followed by a simple element-wise nonlinearity by using gating units. With the gates, the model can decide by itself to which extend to receive, remember and output the incoming information and cell state from the last time-step. Both of these two RNN variants have been shown to perform well in tasks that require capturing long-term dependencies including speech recognition, language modelling, machine translation and etc..

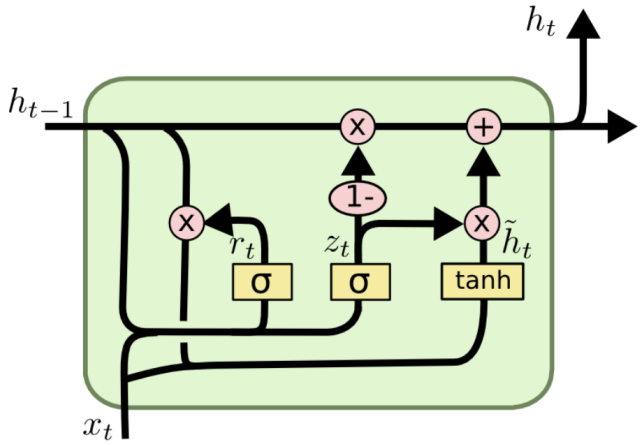

Figure 2.6: The Gated recurrent unit cell.
Source: https://colah.github.io/posts/2015-08-Understanding-LSTMs/

Transformer

Recurrent models typically factor computation along the symbol positions of the input and output sequences. Aligning the positions to steps in computation time, they generate a sequence of hidden state h_t, as a function of the previous hidden state h_{t-1} and the input for position x_t. This inherently sequential nature precludes parallelization within training examples, which becomes critical at longer sequence lengths, as memory constraints limit batching across examples (Vaswani et al., 2017). Vaswani et al. (2017) proposed the Transformer, as the architecture illustrated in Figure 2.7, a model eschewing recurrence and relying entirely on an self-attention mechanism to capture global dependencies between input and output, without using sequence-aligned RNNs or convolution. They proposed a multi-head attention layer that has been fed in with three vectors: the query, keys and values. The output is computed as a weighted sum of the values, where the weight assigned to each value is computed by a compatibility function of the query with corresponding key, as shown in Equation 2.3.

$$Attention(Q, K, V) = softmax(\frac{QK^T}{\sqrt{d_k}})V \tag{2.3}$$

In addition, they found it beneficial to linearly project the queries, keys and values h times with different, learned linear projections to allow the model to jointly attend to information

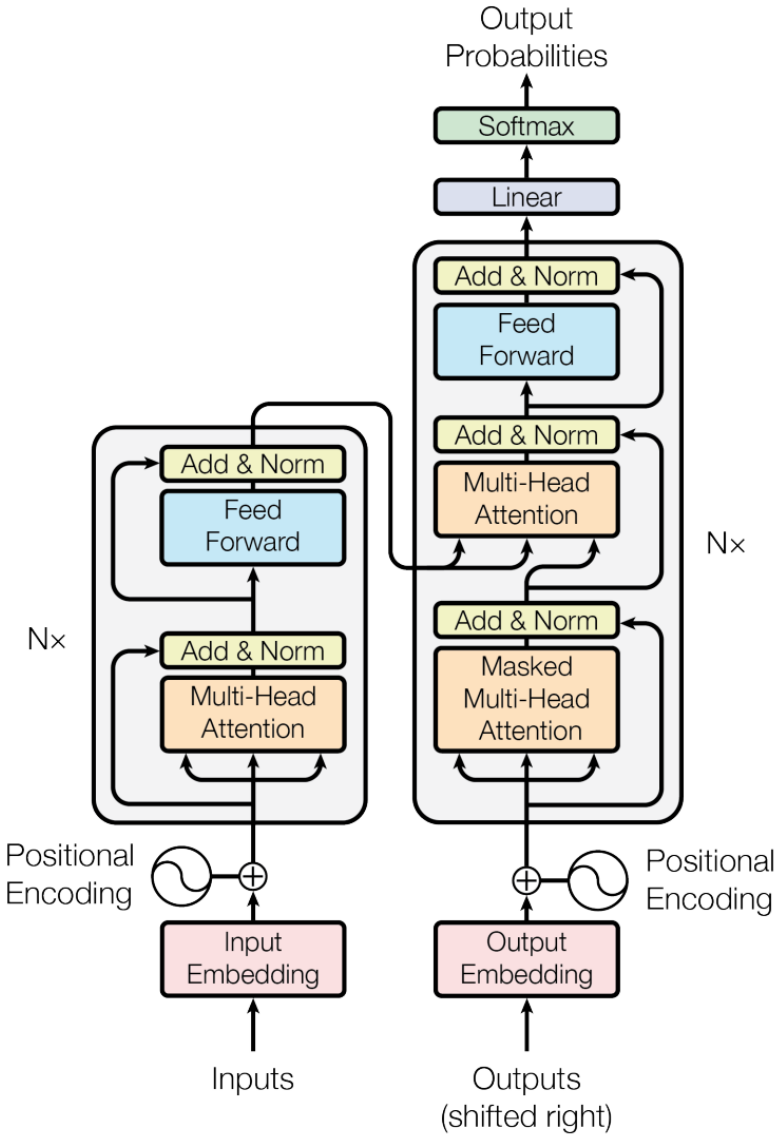

Figure 2.7: The architecture of Transformer from Vaswani et al. (2017).

from different representation subspaces at different positions, which eventually formed the Multi-Head attention, as illustrated in Equation 2.4.

$$\begin{aligned} MultiHead(Q, K, V) &= Concat(head_1, head_2, ..., head_h)W^O \\ where head_i &= Attention(QW_i^Q, KW_i^K, VW_i^V) \end{aligned} \tag{2.4}$$

2.4 Related work

Soricut and Marcu (2003) firstly addressed the task of discourse structure parsing within the same sentences. They proposed two probabilistic models that can be used to identify elementary discourse units and build sentence-level discourse parse trees. They showed how syntactic and lexical information can be exploited in the process of identifying elementary units of discourse and achieved near-human levels of performance on the sophisticated task of deriving sentence-level syntactic trees and discourse segments. A discourse parser requires many components to form a long pipeline. However, the implicit discourse relation classification has been shown to be the main performance bottleneck of an end-to-end parser (Lin et al., 2014). The release of PDTB, the largest available manually annotated corpora of discourse relations, opens the door to supervised machine learning based discourse relation classification. They are roughly categorized into two major types: feature-based and neural network models. Early work addressing discourse relation parsing were trying to classify unmarked discourse relation by training on explicit discourse relations with the marker removed (Marcu and Echihabi, 2002). While this method promised to provide almost unlimited training data, it was shown that explicit relations differ in systematic ways from implicit relations (Asr and Demberg, 2012), so that performance on implicits is very poor when learning on explicits only (Sporleder and Lascarides, 2008). The implicits are difficult and thus is the focus of this thesis.

2.4.1 Feature-based methods

The core idea of feature-based methods is to exploit discriminative features for implicit discourse relations. Pitler et al. (2009) investigated the effectiveness of various features de-

signed to capture lexical and semantic regularities for identifying the sense of implicit discourse relations, including word-pairs, polarity tags, inquirer tags, verb features, modality, context as well as some language modeling features. Results indicated that features developed to capture work polarity, verb classes and orientation, as well as some lexical features are strong indicators for this task. Lin et al. (2009) further introduced contextual, constituent and dependency parse features and achieved an accuracy of 40.2% for 11-way classification, a 14.1% absolute improvements over the baseline. With all these features, Park and Cardie (2012) provided a systematic study of the combinations of them for implicit discourse relation identification and identified feature combinations that optimize F_1-scores for the PDTB to date. They also found that with some other features that are designed to represent the specific aspects for the discourse relation arguments, general features like work pairs may no longer have a role to play for this task. To address the sparsity problem of word pairs, Biran and McKeown (2013) proposed to use relation specific word similarity. They presented a reformulation of the word pair features which replaces the sparse lexical features with dense aggregated score features. Rutherford and Xue (2014) investigated the effects of using Brown clusters as an alternative word representation and analyze the impactful features that arise from Brown cluster pairs. What's more, they studied coreferential patterns in different types of discourse relations in addition to using them to boost the performance of classification.

2.4.2 Neural network methods

With the popularity of using distributed word representations (Bengio et al., 2003; Mikolov et al., 2013), various neural network models have been proposed and proved to be helpful for implicit discourse relation classification. Studies (e.g. Braud and Denis (2015)) have shown that distributed word representations have an advantage in dealing with data sparsity problem that feature-based models have suffered from.

Zhang et al. (2015) proposed a shallow convolutional neural network for implicit discourse relation recognition to alleviate the overfitting problem and help preserve the recognition and generalization ability with the model. Ji and Eisenstein (2014) proposed to get distributed meaning representation for each discourse argument with recurrent neural network. Ji and Eisenstein (2015) have restructured the RNN around a binary syntactic tree.

A tree-structured RNN is also known as recursive neural network, which has been widely used for parsing and sentiment analysis task (Socher et al., 2013; Tai et al., 2015). The hidden state corresponds to an intermediate non-terminal in the phase structure tree. The root node servers as the feature vector for the classification task. In this way, both syntactic and semantic information of the sentence have been preserved and represented by the root vector.

Ji et al. (2016) introduced a latent variable to recurrent neural network and combined RNN language model and discriminative output layer that predicts the discourse relation. This approach worked well in the top-level 4-way classification. Chen et al. (2016) adopted a gated relevance network to capture the semantic interaction between work pairs. Qin et al. (2016a) proposed to use a character-based model to deal with the insufficient training on rare words. With a character-based model, it is easier to encode morphological information and alleviate the rare word problem. Results showed that with the character-enhanced embeddings, the performance on both multi-class and binary classification outperformed most of the existing methods. Qin et al. (2017) introduced an adversarial neural network to exploit the annotated implicit connective by making the model confused about whether the arguments' representations are from explicit or implicit components. With the adversarial training between generator and discriminator, they tried to project both explicit and implicit examples to one unified space, given that the classification is much easier with the informative connective in the explicit instances.

More recently, with the success made by the pre-trained language models, the performances of lots of tasks have been pushed to a new level. Bai and Zhao (2018) combined different grained text representations, including character, subword, word, sentence and sentence pair levels, jointly predicted connective and relations and achieved great improvements on both accuracy and F_1 scores. He et al. (2020) claimed previous methods that primarily encode two arguments separately and extract the specific interaction patterns have failed to fully exploit the annotated relation signal. Instead, they proposed a novel TransS-driven joint learning architecture which translates discourse relation in low-dimensional embedding space and further exploit the semantic features of arguments with multi-level encoders.

However, with the proposed neural network based models become more and more sophis-

ticated, the number of parameters to be learned also boomed. Most of the neural based methods suffer from the insufficient annotated data. Wu et al. (2016) extracted bilingual-constrained synthetic implicit data from a sentence-aligned English-Chinese corpus. Nie et al. (2019) curated a high quality sentence relation task by leveraging explicit discourse relations with dependency parsing and rule-based rubrics. They proposed the DisSent task, which uses the discourse prediction task to train sentence embedding. With the learned embedding for each argument and results showed good generalization performances on lots of tasks.

2.5 Summary

In this chapter, we first give a brief introduction to the basic concepts about discourse relation, including a general idea about the different theories of discourse relation, which is followed by an overview of the largest available manually annotated discourse relation corpus PDTB. In the following chapters, we will mostly focus on the evaluation on the PDTB. The second part of this chapter are some preliminary knowledge regarding neural network models that are most relevant to this dissertation, and also review related work in implicit discourse relation classification. In the upcoming chapter, we will detail the discussion about the problem and risk of using the current conventional data-split settings and our proposed solutions.

Chapter 3

Limited data problem of implicit discourse relation classification

3.1 Introduction

The community most often uses the Penn Discourse Treebank (PDTB) (Prasad et al., 2008) as a resource for implicit discourse relation classification, and has adopted the usual split into training and test data as used for other tasks such as parsing. Because discourse relation annotation is at a higher level than syntactic annotation, this however means that the test set is rather small, and with the amount of alternative features and, more recently, neural network architectures being applied to the problem, we run a serious risk as a community of believing in features that are successful in getting some improvement on the specific test set but don't generalize at all.

Previous studies show that the presence of connectives can greatly help with classification of the relation and can be disambiguated with 0.93 accuracy (4-ways) solely on the discourse relation connectives (Pitler et al., 2008). In implicit relations, no such strong cue is available and the discourse relation instead needs to be inferred based on the two textual arguments.

In recent studies, various classes of features are explored to capture lexical and semantic

regularities for identifying the sense of implicit relations, including linguistically informed features like polarity tags, Levin verb classes, length of verb phrases, language model based features, contextual features, constituent parse features and dependency parse features (Lin et al., 2009; Pitler et al., 2009; Zhou et al., 2010; Zhang et al., 2015; Chen et al., 2016). The most used dataset in this community is the Penn Discourse Treebank, which adds a discourse layer to the Penn Treebank (Marcus et al., 1993). For some of second-level relations in PDTB (a level of granularity that should be much more meaningful to downstream tasks than the four-way distinction), there are only a dozen instances, so that it's important to make maximal use of both the data set for training and testing. The test set that is currently most often used for 11 way classification is section 23 (Lin et al., 2009; Ji and Eisenstein, 2015; Rutherford et al., 2017a), which contains only about 761 implicit relations. This small size implies that a gain of 1 percentage point in accuracy corresponds to just classifying an additional 7-8 instances correctly, which makes it risky to draw conclusions solely relying on the results of the limited test data.

This chapter therefore aims to demonstrate the degree to which conclusions about the effectiveness of including certain features would depend on whether one evaluates on the standard test section only, or performs cross validation on the whole dataset for second-level discourse relation classification. The model that we use is a neural network that takes the words occurring in the relation arguments as input, as well as traditional features mentioned above, to make comparisons with most-used section splits. To the best of our knowledge, this was the first work that systematically evaluates the effect of the train/validation/test splits for the implicit discourse relation classification task on PDTB at the time of the research. In this chapter, We report the classification performances (accuracy) on random and conventional splits among the whole sections.

As a model, we use a neural network that also includes some of the surface features that have been shown to be successful in previous work, which was as well competitive with the state of the art models at the time of the study. The experiments here are exemplary of what kind of conclusions we would draw from the cross validation vs. from the usual train-test split. We find that results are quite different in the different splits of dataset, which we think is a strong indication that cross validation is important to adopt as a standard practice

for the implicit discourse relation classification community. We view cross validation as an important method in case other unseen datasets are not available (note that at least for English, new datasets have recently been made available as part of the shared task (Xue et al., 2015, 2016), as well as SPICE (Rehbein et al., 2016)).

The model we use in this chapter is most closely related to the neural network model proposed in Rutherford et al. (2017a). The model also has access to the traditional features, which are concatenated to the neural representations of the arguments in the output layer. In order to simulate what conclusions we would be drawing from comparing the contributions of the handcrafted surface features, we calculate accuracy for each of the hand-crafted features.

3.2 Corpora and conventional settings

The Penn Discourse Treebank (PDTB) We use the Penn Discourse Treebank (Prasad et al., 2008), the largest available manually annotated corpora of discourse on top of one million word tokens from the Wall Street Journal (WSJ). The PDTB provides annotations for explicit and implicit discourse relations. By definition, an explicit relation contains an explicit discourse connective while the implicit one does not. The PDTB provides a three level hierarchy of relation tags for its annotation. Previous work in this task has been done over two schemes of evaluation: first-level 4-ways classification (Pitler et al., 2009; Rutherford and Xue, 2014; Chen et al., 2016), second-level 11-ways classification Lin et al. (2009); Ji and Eisenstein (2015). The distribution of second-level relations in PDTB is illustrated in Table 3.1.

We follow the preprocessing method in Lin et al. (2009) and Rutherford et al. (2017a). If the instance is annotated with two relations, we use the first one shown up, and remove those relations with too few instances (less than 5, same as previous work). We treat section 2-21 as training set, section 22 as development set and section 23 as test set for our results reported as "most-used split". In order to investigate whether the results with benefits from including a certain feature to the model are stable, we conduct 10-fold cross-validation on the whole corpus including sections 0-24. Note that we here included also the validation section for

Relation	Most-used Split				Cross Validation *	
	Train		Test		Train	Test
Temporal.Asynchronous	542	(4.25%)	12	(1.58%)	583	65
Temporal.Synchrony	150	(1.18%)	5	(0.66%)	155	18
Contingency.Cause	3259	(25.53%)	193	(25.36%)	3581	398
Contingency.Pragmatic cause	55	(0.43%)	5	(0.66%)	61	7
Comparison.Contrast	1600	(12.54%)	126	(16.56%)	1843	205
Comparison.Concession	189	(1.48%)	5	(0.66%)	194	22
Expansion.Conjunction	2869	(22.48%)	116	(15.24%)	3075	342
Expansion.Instantiation	1130	(8.85%)	69	(9.07%)	1254	140
Expansion.Restatement	2481	(19.44%)	190	(24.97%)	2792	311
Expansion.Alternative	151	(1.18%)	15	(1.97%)	160	18
Expansion.List	337	(2.64%)	25	(3.29%)	347	39
Total	12763		761		14045	1565

* Numbers are averaged over different folds

Table 3.1: The distribution of training and test sets in Most-used Split and Cross Validation on level 2 relations in PDTB. Five types that have only very few training instances are removed.

our experiments, to have maximal data for our demonstration of variability between folds. For best practice when testing new models, we instead recommend to keep the validation set completely separate and do cross-validation for the remaining data. Also note that you might want to choose repeated cross-validation (which simply repeats the cross-validation step several times with the data divided up into different folds) as an alternative to simple cross-validation performed here. For a more in-detail discussion of cross validation methods, see Kim (2009); Bengio and Grandvalet (2005).

In Table 3.1, we can see that the relations' proportions are quite different on the training and test set of the most-used split setting. For instance, temporal relations are under-represented which may lead to a misestimation of the usefulness of features that are relevant for classifying temporal relations. For our cross validation experiments, we evenly divide all the instances in section 0-24 into 10 balanced folds[1]. The proportions of each class in the training and testing set are identical. With the same distribution of each class, we here avoid having an unbalanced number of instances per class among training and testing set.

[1]While we here chose balanced distributions, other designs of splitting up the data into folds such that different folds have organically different distributions of classes can alternatively be argued for, on the basis of more accurately representing new in-domain data distributions.

3.3 Approach

3.3.1 Overview of the model

The task is to predict the implicit discourse relation given the two arguments of an implicit instance. As a label set, we use 11-way distinction as proposed in Lin et al. (2009) and Ji and Eisenstein (2015). Word Embeddings are trained with the Skip-gram architecture in *Word2Vec* (Mikolov et al., 2013), which is able to capture semantic and syntactic patterns with an unsupervised method, on the training sections of WSJ data.

Our model is illustrated in Figure 3.1. Each word is represented as a vector, which is found through a look-up word embedding. Then we get the representations of argument 1 and argument 2 separately after transforming semantic word vectors into distributed continuous-value features by LSTM recurrent neural network. With concatenating feature vector and the instance's representation, we classify it with a softmax layer and output its predicted label.

Implementation All the models are implemented in Keras[2], which runs on top of Theano. The architecture of the model we use is illustrated in Figure 3.1. Regarding the initialization, regularization and optimization, we follow all the settings in Rutherford et al. (2017a). We employ cross-entropy as our cost function, Adagrad as the optimization algorithm, initialized all the weights in the model with uniform random and set dropout layers after the embedding and output layer with a drop rate of 0.2 and 0.5 respectively.

3.3.2 Features

For the sake of our cross-validation argument, we choose five kinds of most popular features in discourse relation classification, namely *Inquirer Tags* (semantic classification tags), *Brown Clusters*, *Verb* features, *Levin classes* and *Modality*. Features that include word pairs directly are not included here, because we assume these have already been represented directly in the neural network by the concatenated representations of the arguments.

Inquirer Tags: Negated and non-negated fine-grained semantic classification tags for the

[2]https://keras.io/

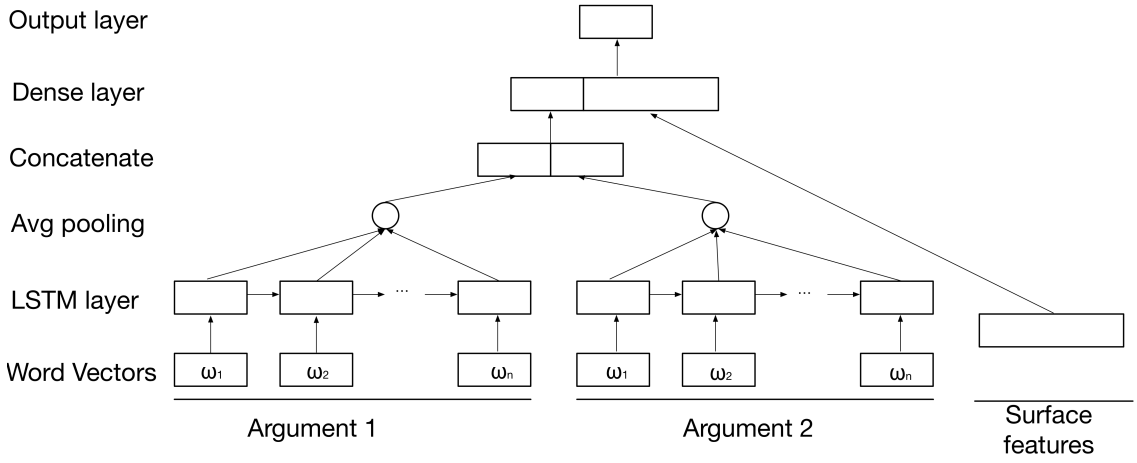

Figure 3.1: Long Short-Term Memory Model with surface features.

verbs in each arguments. They are drawn from the General Inquirer Lexicon (Stone et al., 1966).

Brown Cluster: Brown Clustering algorithm (Brown et al., 1992) induces a hierarchy of words in a large unannotated corpus based on word co-occurrences within the window. Thus every word has a unique hierarchical word cluster.

Verbs: Including Part-of-Speech of the main verbs, the Average Lengths of Verb Phrase, count of verbs from Arg1 and Arg2 belonging to the same Levin Class (Levin, 1993).

Modality: Modal words, which are often used to express conditional statements. We include a feature for the presence and absence of modal words in argument 1 and 2.

3.4 Results

We test five frequently-used surface features with our model. Results are shown in Table 3.2. We can see that our implemented model is comparable with the state of the art models at the time of the study. Our main point here is however not to argue that we outperform any particular model, but rather we would like to discuss what conclusions we would be drawing from adding surface features to our neural network model if using the standard test set vs. doing cross validation.

In Table 3.2, “No additional surface features” means that there is no surface feature in the

	Models	Most-used Split	Cross Validation
	Most common class	25.36	25.59
	Lin et al. (2009)	40.20	-[1]
	Ji and Eisenstein (2015) (surface features only)	40.66	-
	Rutherford et al. (2017a)	39.56	-
Neural Network	No additional surface features	37.68	34.44 ($\pm$1.37)
Neural Network	Inquirer Tags	40.46	33.58 ($\pm$1.36) (2+,8-)
Neural Network	BrownCluster	38.77	33.83 ($\pm$1.59) (3+,7-)
Neural Network	Levin Class	40.92	34.17 ($\pm$1.48) (4+,6-)
Neural Network	Verbs	40.21	34.26 ($\pm$1.22) (5+,5-)
Neural Network	Modality	40.82	37.65 ($\pm$1.83) (6+,4-)
Neural Network	All Features above	38.56	35.90 ($\pm$1.32) (2+,8-)

[1] "-" means no result currently.

Table 3.2: Performance comparison of different features in Most-used Split and Cross Validation on second-level relations. Numbers for cross validation indicate the mean accuracy across folds, the standard deviation, and the number of folds that show better vs. worse performance when including the feature.

input. It is easy to find that all the performances on the most-split settings are improved by the surface features with 1-3%. But it's not in the same case with the cross-validation setting. For instance, the Levin class helps improve the accuracy with 3.2% on the conventional setting but lowers the accuracy on the cross-validation.

For each cross validation with different features, the separation into train and test sets are identical. We can see that the performances on Most-used Split section is generally 3-7% better than the results for the rest of the corpus. While we would also conclude from our model when evaluated on the standard test set that each of these features contribute some useful information, we can also see that it comes to very different conclusions if actually running the cross-validation experiment.

The cross validation is primarily a way of measuring the predictive performance of a model. With such a small test set, improvements on the classification could be the results of many factors. For instance, take a look at the effectiveness of including Inquirer Tags: these lead to an increase in performance by 2.8% in Most-used Split, but actually only helped on two out of 10-fold in the cross-validation set, overall leading to a small decrease in performance of the classifier. Similarly, the verb features seem to indicate a substantial improvement in relation classification accuracy on the standard test set, but there is no effect at all across the

folds.

It is easy to understand from Table 3.2 that performances on most-used split sections are improved by the surface features. To verify if it is the same on the cross validation, we employ the Student's t-test to do the significance test. The p-value for the performances with inquirer tags, brown cluster, verbs and modality are 0.14, 0.25, $2.6e^{-9}$, $0.7e^{-3}$, which mean that we cannot say that all the features are useful to improve the classification accuracy significantly, compared with the No Feature one. That may be due to the different distribution of relations or the features indeed cannot be generalized to the whole dataset.

Nonetheless, we see that cross validation is not the only way to validate the problem of different distribution in a relative small dataset. Other works, such as Berg-Kirkpatrick et al. (2012) strongly recommend significance testing to validate metric gains in natural language processing tasks, even though the relationship between metric gain and statistical significance is complex. We observe that recent papers in discourse relation parsing do not always perform significance testing, and if they do report significance, then oftentimes they do not report the test that was used. We would here like to argue in favour of significance testing with cross validation, as opposed to boot strapping methods that only use the standard test set. Due to the larger amount of data, calculating significance based on the cross validation will give us substantially better estimates about the robustness of our results, because it can quantify more exactly the amount of variation with respect to transferring to a new (in-domain) dataset.

3.5 Conclusion and discussion

In this chapter, we have argued that the standard test section of the PDTB is too small to draw conclusions upon, about whether a feature is generally useful or not, especially when using a larger label set as is the case in recent work using second level labels.

While these ideas are far from new and apply also to other NLP tasks with small evaluation sets, we think it is important to discuss this issue, as recent work in the field of discourse relation analysis has mostly ignored the issue of small test set sizes in the PDTB.

Our experiments support our claim by showing that features that may look like they improve

performance on the 11-way classification on the standard test set, do not always show a consistent improvement when the training / testing was split up differently. This means that we run a large risk of drawing incorrect conclusions about which features are helpful if we only stick to our small standard test set for evaluation.

3.6 Summary

In this chapter, we show that it is risky to draw conclusions about features or models with a limited size of test set. However, cross validation is only an important method and concession without new annotated dataset.

Having cheaper, more reliable and automatically labeled data would be much beneficial for this task, especially for the various neural network models proposed recently. With more data, how to have better understanding of how arguments relate to one another and to have better semantic representations are also crucial for the task. However, having good encoding only does part of the job, a good implicit discourse relation classifier should also be competent in being able to encode discourse expectation and learn typical events, causes, consequences etc. for all kind of events. In the next chapters, we address the data bottleneck problem by proposing different methods to solve each of the problems separately, including introducing a new pipeline to get more annotated implicit discourse relation instances with the help of explicitated connectives between English and French translations, proposing a sequence-to-sequence model to encode the process of explicitation, trying to figure out the importance of having correct next sentences and also shifting the model across domains.

Chapter 4

Explicitation of Implicit Discourse Relation between English and French

4.1 Introduction

Implicit relation classification is very challenging and represents a bottleneck of the entire discourse parsing system.

In recent studies, lots of methods are proposed to directly infer underlying relations, ranging from various classes of features, to the end-to-end neural models.

Early methods have focused on designing various features to overcome data sparsity and more effectively identify relevant concepts in the two discourse relational arguments. (Lin et al., 2009; Zhou et al., 2010; Biran and McKeown, 2013; Park and Cardie, 2012; Rutherford and Xue, 2014), while recent efforts use distributed representations with neural network architectures (Zhang et al., 2015; Ji and Eisenstein, 2015; Ji et al., 2016; Chen et al., 2016; Qin et al., 2016a, 2017). Both streams of methods suffer from insufficient annotated data (Wang et al., 2015), since the Penn Discourse Treebank (PDTB) (Prasad et al., 2008), which is the discourse annotated resource mostly used by the community, consists of just 12763 implicit instances in the usual training set and 761 relations in the test set. Some second-

level relations only have about a dozen instances.

With such limited amount of data, in the last chapter, we argue that we would run a serious risk as the community believes in features that are successful in getting some improvements on the specific test set but don't generalize well. Having more labeled data for training would lower the risk and generalize the trained model to have more stable and reliable performances even with a large label set, like the PDTB label set used in the implicit discourse relation classification task. It is therefore crucial to obtain extra data for machine learning methods.

However, manually annotating implicit discourse relation is a very difficult, time-consuming and expensive task. In this chapter, we propose a simple approach to automatically extract samples of implicit discourse relations from parallel corpus via back-translation: Our approach is motivated by the fact that humans sometimes omit connectives during translation (*implicitation*), or insert connectives not originally present in the source text (*explicitation*) (Laali and Kosseim, 2014; Koppel and Ordan, 2011; Cartoni et al., 2011; Hoek and Zufferey, 2015; Zufferey, 2016). When explicitating an implicit relation, the human translator is, in other words, disambiguating the source implicit relation with an explicit DC in the target language. This chapter focuses on the former case, but it also can be applied in reverse, to find less explicit as well.

The contribution of this chapter is twofold: Firstly, we propose a pipeline to automatically label English implicit discourse relation samples based on explicitation of DCs in human translation, which is the target side of a parallel corpus. Secondly, we show that the extra instances mined by the proposed method improve the performance of a standard neural classifier by a large margin, when evaluated on the PDTB 2.0 benchmark test set as well as by cross-validation which is advocated in the last chapter.

4.2 Related work

Early work addressing discourse relation parsing were trying to classify unmarked discourse relations by training on explicit discourse relations with the marker been removed (Marcu and Echihabi, 2002). While this method promised to provide almost unlimited training

data, it was shown that explicit relations differ in systematic ways from implicit relations (Asr and Demberg, 2012), so that performance on implicits is very poor when learning on explicits only (Sporleder and Lascarides, 2008).

The release of PDTB (Prasad et al., 2008), the largest available corpus which annotates implicit examples, lead to substantial improvements in classification of implicit relations, and spurred a variety of approaches to the task, including feature-based methods (Pitler et al., 2009; Lin et al., 2009; Park and Cardie, 2012; Biran and McKeown, 2013; Rutherford and Xue, 2014) and neural network models (Zhang et al., 2015; Ji and Eisenstein, 2015; Ji et al., 2016; Chen et al., 2016; Qin et al., 2016a, 2017). However, the limited size of the annotated corpus, in combination with the difficulty of the task of inferring the type of relation between given text spans, presents a problem both in training (Rutherford et al. (2017c) find that a simple feed-forward architecture can outperform more complex architectures, and argue that the larger number of parameters can not be estimated adequately on the small amount of training data) and testing (In the Chapter 3, we show that results on the standard test set are not reliable due to the small set of just 761 relations).

Data extension has therefore been a longstanding goal in discourse relation classification. The main idea has been to select explicit discourse instances that are similar to implicit ones to add to the training set. Wang et al. (2012) proposed to differentiate typical and atypical examples for each discourse relation, and augment training data for implicits only by typical explicits. In a similar vein, Rutherford and Xue (2015) proposed criteria for selecting among explicitly marked relations ones that contain discourse connectives which can be omitted without changing the interpretation of the discourse. These relations are then added to the implicit instances in training.

On the other hand, Lan et al. (2013) presented multi-task learning based systems, which in addition to the main implicit relation classification task, contain the task of predicting previously removed connectives for explicit relations, and profit from shared representations between the tasks. Similarly, Hernault et al. (2010) observes features that occur in both implicit and explicit discourse relations, and exploit such feature co-occurrence to extend the features for classifying implicits using explicitly marked relations. Mihăilă and Ananiadou (2014) and Hidey and McKeown (2016) proposed semi-supervised learning and self-

learning methods to improve recognition of patterns that typically signal causal discourse relations.

The approach proposed here differs from previous approaches, because we extend our training data only by originally implicit relations, and obtain the label through the disambiguation that sometimes happens in human translation.

Parallel corpora have been exploited as a resource of discourse relation data in previous work but have mostly been used with goals different from ours: Cartoni et al. (2013) and Meyer et al. (2015) used parallel corpora to label and disambiguate discourse connectives in the target language based on explicitly marked English relations, in order to help machine translation. A second application has been to project discourse annotation from English onto other languages through parallel corpora, in order to construct discourse annotated resources for the target language (Versley, 2010; Zhou et al., 2012; Laali and Kosseim, 2014).

The approach that is in spirit most similar to ours is by Wu et al. (2016), who extracted bilingual-constrained synthetic implicit data from a sentence-aligned English-Chinese corpus and got improvements by incorporating these data via a multi-task neural network on the 4-way classification.

4.3 System overview

Our proposed method aims at sentence pairs in the parallel corpora where an *implicit* discourse relations on the source English side has been translated by human translators into an explicitly marked relation on the target side. The inserted connective hence disambiguates the originally implicit relation, and the discourse relation can be classified with confidence (under the assumption that the same discourse relation holds in the original source text).

The pipeline of our approach is detailed in below steps.

1. The target side of a sentence-aligned parallel corpus, with English as the source text, is back-translated to English using a pre-trained machine translation system.
2. An end-to-end discourse relation parser for English is run on both the source side and the back-translated target side. The parser will output a list of explicit and implicit relations, including the relation sense and argument spans of each relation.

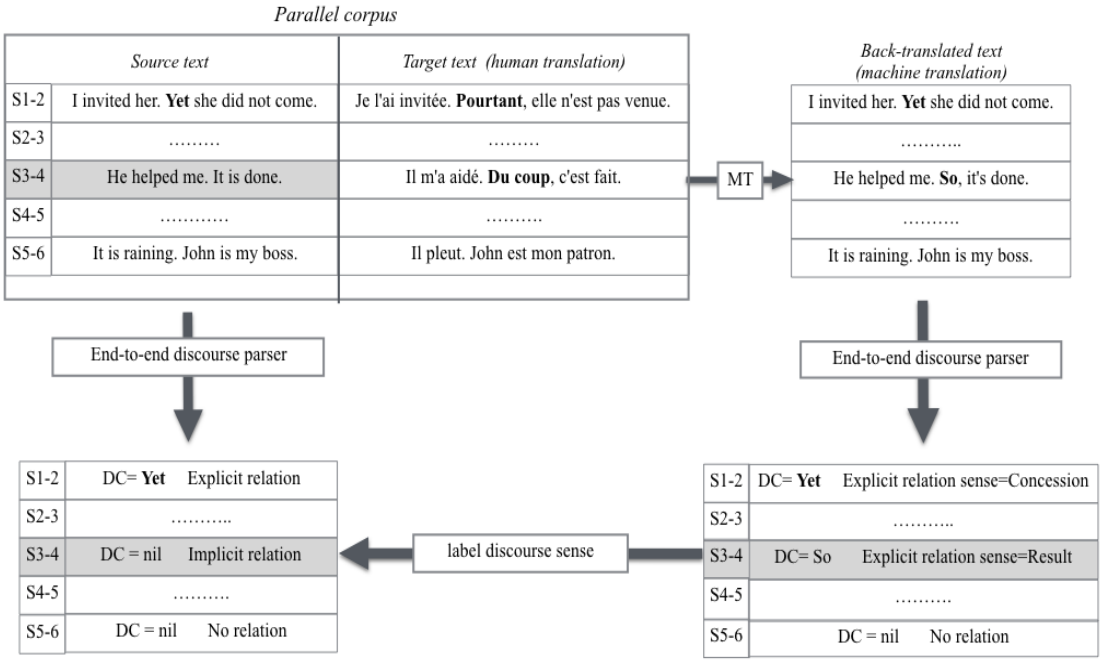

Figure 4.1: Pipeline showing how an implicit discourse relation sample, sentence pair 3-4, is extracted and labeled using a parallel corpus.

3. Implicit-to-explicit discourse relation alignments are identified according to the output of the end-to-end parser. Implicit relations in the PDTB are only ever annotated between consecutive sentences. Therefore, we specifically extract pairs of consecutive sentences on the source English side:

    - that are identified as the *Arg1* and *Arg2* of an *implicit* discourse relation[1];
    - whose corresponding back-translated target sentences are identified as the *Arg1* and *Arg2* of an *explicit* relation;
    - that are not part of the *Arg1* or *Arg2* of any other discourse relations[2].

4. Label the source English implicit relation with the relation class of the explicit relation in back-translated target text. The two consecutive sentences are marked as *Arg1* and *Arg2* respectively.

Figure 4.1 illustrates the pipeline of our approach, which takes an English-to-French parallel corpus as input and outputs a list of implicit discourse relations, each containing two

[1] Relations signaled by *Alternative Lexicalization* are counted as implicit relations and extracted as samples. However, *NoRel* and *EntRel* are excluded.

[2] This restriction avoids mis-alignment of relations between source and target texts.

arguments from the source English text and a relation class according to the back-translated French DC.

We then compare the performance of a neural implicit discourse relation classifier trained with the annotated implicit relation samples in PDTB alone and also with the extra training samples mined from the parallel corpus. The classifier performance is evaluated on the standard PDTB implicit relation test set and by cross-validation.

4.3.1 Advantages of using back-translation

In the proposed method, we disambiguate implicit relations according to the explicitated translation. Instead of directly classifying the explicit relation in the target language, we back-translate the target text to the source language by machine translation (MT) because:

- Discourse parsers on low-resource languages do not perform well, or are even not available.
- Different languages have different sets of discourse relation classes defined. By the means of back-translation, we can use an English discourse parser on the target text, and thus label the implicit relations with the same set of relation labels defined for English.
- The quality of the MT system has limited impact on our approach. Since the DC tokens are powerful features to disambiguate an explicit relation, limited contextual features are required. We just need correct translation of the explicit DC tokens, irrespective of word order and the rest of the translation.

4.3.2 Inter-sentential and intra-sentential relations

Only inter-sentential implicit relations are annotated in the PDTB, due to time and resource constraints (Prasad et al., 2008). However, this does not mean that implicit relations only hold between consecutive sentences.

We decided to extract intra-sentential relation samples from the parallel corpus based on two motivations: Firstly, we hypothesize that intra-sentential implicit relations share similar features as inter-sentential ones. Including both types may hence increase dataset size.

In fact, we will see in the experiment results that intra-sentential training samples largely improve classification of implicit relations, even though the test data from PDTB contains inter-sentential samples only. An analysis on what we learn from the intra-sentential samples is presented in Section 4.5.2.

Secondly, intra-sentential relations can potentially be identified with higher reliability: Parallel corpora are typically sentence-aligned. This makes it a lot easier to extract sentences that are detected by the end-to-end discourse relation parser as explicit in the (back-)translation target side but not on the original source side, without needing to worry about whether any sentences in the dataset were removed or the order changed during preprocessing (which would be detrimental for detecting intra-sentential relations).

4.3.3 Argument spans

It is possible but not entirely trivial to determine the argument spans of the discourse relations labeled with the back-translation method. In this chapter, we chose a neural network model that concatenates the *Arg1* and *Arg2* representations (see Section 4.4.4), so that determining exact text spans of Arg1 and Arg2 was not necessary. We are not the first one to do like this, in the work by Rönnqvist et al. (2017), they modeled the *Arg1-Arg2* pairs as a joint sequence and did not compute intermediate representations of arguments separately, to make it more generally flexible in modeling discourse units and easily extend to additional contexts.

4.4 Experiments

4.4.1 Data

Parallel Corpora The corpora used for the extraction of implicit discourse relation samples are publicly available bilingual English-French parallel datasets compiled by (Rabinovich et al., 2015).[3] They consist of European parliamentary proceedings, literary works and the Hansard corpus – genres that are different from the PDTB, because we want to expand the

[3] All corpora are available at `http://cl.haifa.ac.il/projects/translationese/`

diversity of discourse relation samples available in the PDTB. These corpora contain a total of ~ 1.9M sentence pairs with an average of 22.7 words per English sentence. Each corpus contains an originally written part in English (used as target for the MT system) and its corresponding human translation in French (used as source). We use the same corpora to train the French–English MT system (Section 4.4.2), to back-translate the French side into English and to extract additional discourse training data.

The Penn Discourse Treebank (PDTB) We use the Penn Discourse Treebank 2.0 (Prasad et al., 2008) for the training and testing of the implicit discourse relation classifier. PDTB is the largest available manually annotated corpus of explicit and implicit discourse relations based on one million word tokens from the Wall Street Journal. Each discourse relation is annotated with at most two senses from a three-level hierarchy of discourse relations: Class, Type and Subtype. The first level roughly categorizes the relations into four major classes, each of which is further categorized in to more distinct relation types. Conventionally, discourse relation classifiers are either evaluated by the accuracy of the first-level 4-way classification(Pitler et al., 2009; Rutherford and Xue, 2014; Chen et al., 2016), or the second-level 11-way classification (Lin et al., 2009; Ji and Eisenstein, 2015; Qin et al., 2016a, 2017).

4.4.2 Machine translation system

We train an MT system to back-translate the target side of the parallel corpus to English. To produce the highest-quality back-translation, we use a neural MT system trained on the same parallel corpus. The system is implemented by Open-source Neural Machine Translation (OpenNMT), as shown in Figure 4.2. (Klein et al., 2017).

Source words are first mapped to word vectors and then fed into a recurrent neural network. At each target time step, *attention* is applied over the source RNN and combined with the current hidden state to produce a prediction of the next word, and this prediction would be fed back into the target RNN.

We evaluate the MT system on *newstest2014* and *newsdiscusstest2015*, reaching 24.63 and 22.58 BLEU respectively. The French side of the training data back-translated into English

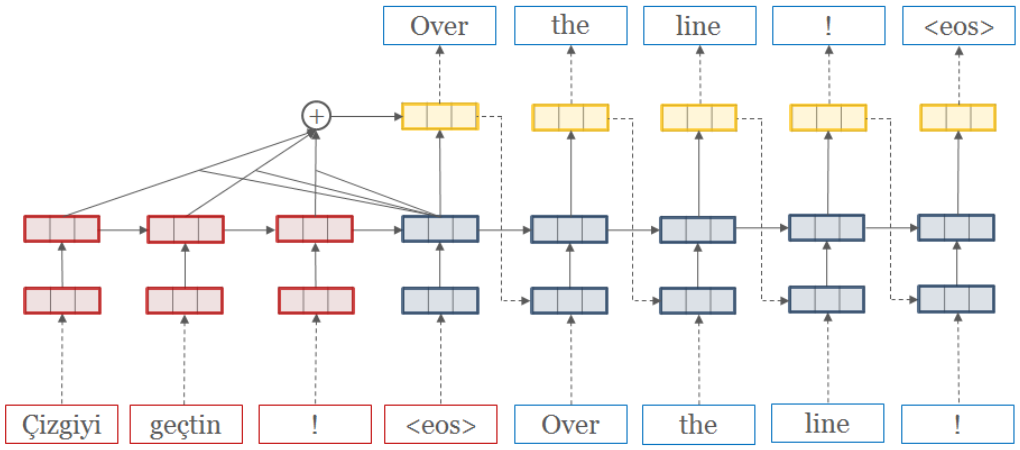

Figure 4.2: Schematic view of neural machine translation (NMT).

is evaluated against the originally written English source, leading to a BLEU score of 34.17.[4] The evaluation of the back-translated corpus indicates that the source text is not exactly reproduced. Critically, we assume that the MT system preserves the explicitness of the target DCs, instead of explicitating or implicitating DCs as in the human translation.

4.4.3 End-to-end discourse parser

We employ the PDTB-style End-to-End Discourse Parser (Lin et al., 2014) to identify and classify the explicit instances from the back-translated English sentences. It achieved about 87% F1 score for explicit relations on level-2 types, even higher than human agreement of 84%. The accuracy on explicit DC identification is 96%.

On the source side, the end-to-end parser is applied to pick implicit relations from other types of relations, i.e. explicit relations or *no relation*, in order to extract implicit-to-explicit DC translation from the parallel corpus[5]. On the back-translation, the end-to-end parser is applied to identify only explicitly marked discourse relations.

[4]Case sensitive BLEU implemented in *mteval-v13a.pl.* Test sets available at `http://www.statmt.org/wmt15/translation-task.html`

[5]The non-explicit sense classification module of this parser is thus not used in the proposed method.

4.4.4 Implicit relation classification model

We use a Bidirectional Long Short-Term Memory (LSTM) network as the implicit relation classification model to evaluate the samples extracted by the proposed method. This architecture inspects both left and right contextual information and has been proven effective in relation classification (Zhou et al., 2016; Rönnqvist et al., 2017). The reasons why we choose this model come from the following two sides. Firstly, bidirectional LSTM network combines forward and backward sequence representations, which could better capture dependencies between parts of the input sequence by inspection of both left and right-hand-side context at each time step. Secondly, it is very easy to implement and also effective in multiple NLP tasks (Rönnqvist et al., 2017; Zhou et al., 2016).

The model is illustrated in Figure 4.3, where each word from the two discourse relational arguments is represented as a vector, which is found through a look-up word embedding. Given the word representations [$\mathbf{w}_1$,$\mathbf{w}_2$,...,$\mathbf{w}_n$] as the input sequence, an LSTM computes the state sequence [$\mathbf{h}_1$,$\mathbf{h}_2$,...,$\mathbf{h}_n$] with the following equations:

$$\begin{aligned} \begin{bmatrix} i_t \\ f_t \\ o_t \\ \hat{c}_t \end{bmatrix} &= \begin{bmatrix} \sigma \\ \sigma \\ \sigma \\ \tanh \end{bmatrix} W \cdot [h_{t-1}, x_t] \\ c_t &= f_t \odot c_{t-1} + i_t \odot \hat{c}_t \\ h_t &= o_t \odot \tanh(c_t) \end{aligned} \tag{4.1}$$

The forward and backward LSTM layers traverse the sequence $\boldsymbol{e}_i$, producing sequences of vectors $\boldsymbol{h}_{if}$ and $\boldsymbol{h}_{ib}$ respectively, which are summed together in the coming sum layer.

Following the preprocessing method in Lin et al. (2009), relations with too few instances (*Contingency.Condition, Pragmatic Condition; Comparison.Pragmatic Contrast, Pragmatic Concession; Expansion.Exception*) are removed during training and evaluation, resulting in 11 types of relations. Among instances annotated with two relation senses, we only use the first sense.

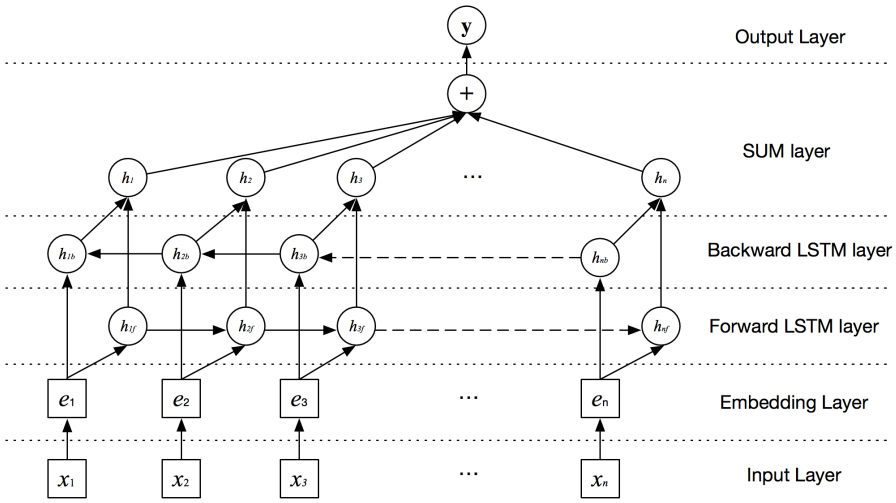

Figure 4.3: The bidirectional LSTM Network for the task of implicit discourse relation classification.

Relation	intra-	inter-	Total
explicit → explicit	199,047	111,090	310,137
explicit → implicit	101,381	29,964	131,345
implicit → explicit	77,228	25,086	102,314

[1] "→" means from source to target side.

Table 4.1: Numbers of intra/inter-sentence samples extracted from parallel corpora.

The model is implemented in Keras[6], which is capable of running on top of Theano. We use word embeddings of 300 dimensions, which are trained on the original English side of the parallel corpora as well as PDTB with the Skip-gram architecture in *Word2Vec* Mikolov et al. (2013). We initialize the weights with uniform random distribution; use standard cross-entropy as our loss function; employ Adagrad as the optimization algorithm of choice and set dropout layers after the embedding layer and output layer with a drop rate of 0.2 and 0.5 respectively. Each LSTM has a vector dimension of 300, matching the embedding size.

We split the PDTB data and evaluate the classifier in two settings. Firstly, we adopt the standard PDTB splitting convention, where section 2-21, 22, and 23 are used as train, validation and test sets respectively (Lin et al., 2009). Secondly, we conduct 10-fold cross validation on the whole corpus including sections 0-24, as advocated in Shi and Demberg (2017). And extra samples are only added into training folds in the CV setting, which means that testing fold consists of instances from PDTB only. Models trained with and without extra samples we extracted, on top of the PDTB data, are compared.

4.5 Distribution of additional instances

In total, $102,314$ implicit discourse relation samples are extracted, of which $25,086$ are inter-sentential relations and $77,228$ are intra-sentential. Inter-sentential relations are much less abundant because stricter screening strategy is applied (the end of point 3 in Chapter 4.3). From Table 4.1 we can also see that majority of DCs in the source side have been translated into the target side explicitly.

Figure 4.4 compares the distribution of relation senses among the annotated implicit relations in the PDTB and our extracted samples. The relation distribution generally corresponds to the distribution in PDTB, but some relations, such as *Temporal* and *Contingency.Condition*, are particularly numerous in the intra-sentential samples.

[6]https://keras.io/

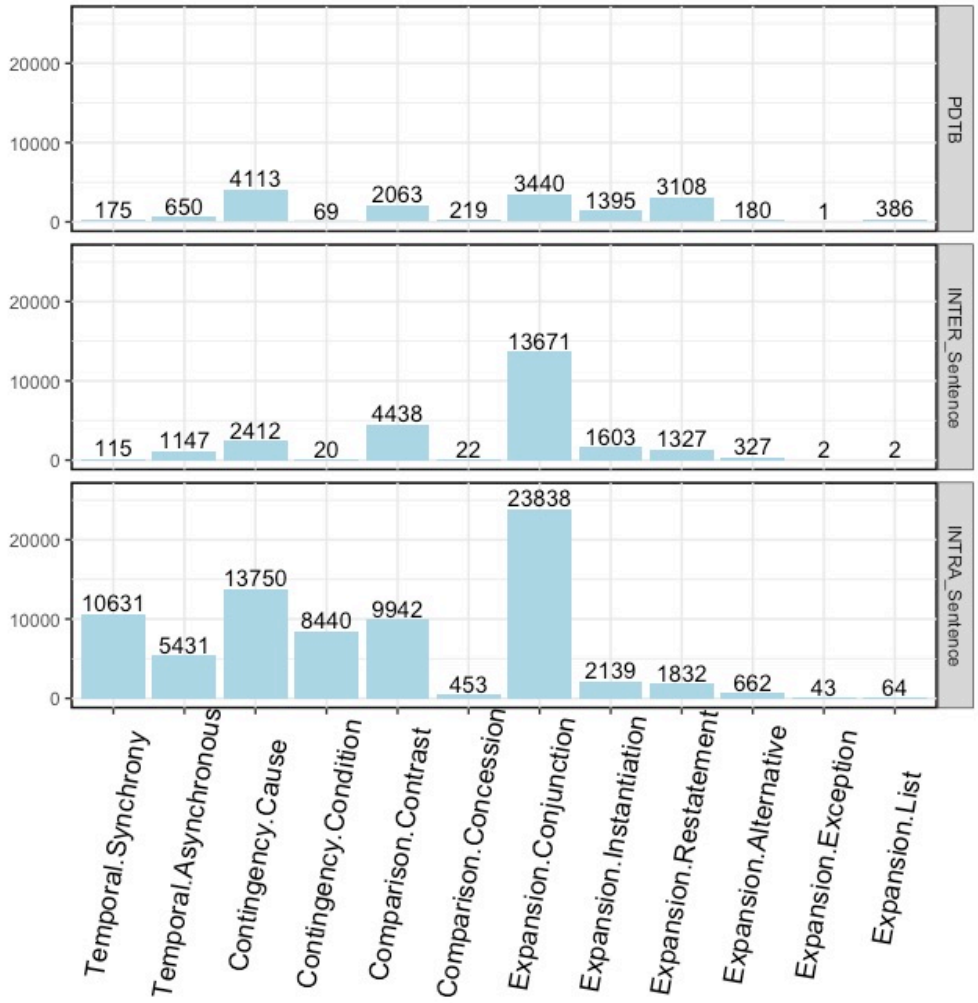

Figure 4.4: Relation sense distribution of implicit relations in PDTB and the extra intra- and inter-sentence samples

4.5.1 Experimental results

We compare our model with current state-of-the-art models that were evaluated under the same setting (11-way classification, PDTB section 23 as test set) (Qin et al., 2016a, 2017; Rutherford et al., 2017c), as well as a model based on linguistic features (Lin et al., 2009) that uses this setting for evaluation.

Qin et al. (2017) developed an adversarial model, which consists of two CNNs in which arguments are represented separately, a four-layer Perceptron and a dense layer for classification, to enable an adaptive imitation scheme through competition between the implicit network and a rival feature discriminator. Our model substantially differs from that setup, as it uses a much simpler network architecture and represents the two discourse relation arguments

Models		PDTB Test Set	Cross Validation
Most common class		25.36	25.59
Lin et al. (2009)		40.20	-
Ji and Eisenstein (2015)	Surface features only	40.66	-
	+ Entity semantics	44.59	-
Qin et al. (2016a)		43.81	-
Qin et al. (2017)		44.65	-
Rutherford et al. (2017c)		39.56	-
Shi and Demberg (2017) (no surface features)		37.68	34.44
Ours	PDTB only	34.32	30.01
	PDTB + inter-sentential samples	42.29	34.14
	PDTB + intra-sentential samples	44.29	35.08
	PDTB + all samples	**45.50**	**37.84**

[1] "-" means no result currently.

Table 4.2: Accuracy of 11-way classification of implicit discourse relations on PDTB test set and by cross validation.

jointly, i.e. without knowledge of the arguments' spans. We can see that our baseline model performs substantially less well than the state of the art, and also less well than our results in Chapter 3, which also uses an LSTM but represents discourse relational arguments separately. As adding training data can be expected to be largely orthogonal to the choice of classification model, we are here most interested in seeing whether adding the new instances improves over the baseline model with identical architecture.

Table 4.2 shows that including the extra inter- and intra-sentential instances leads to very substantial improvements in classification accuracy. Using the additional data, our method not only improves performance by 11%-points on the PDTB test set compared to training on the PDTB implicit relations only, but also outperforms much more complex neural network models (Qin et al., 2016a, 2017) on this task.

The evaluation using cross-validation (around 8% point improvement over the baseline) furthermore shows that the obtained improvements do not only hold for the PDTB standard test set but also are stable across the whole PDTB data. These results strongly support the effectiveness of the implicit relation samples mined from parallel texts.

The accuracies reported for our models are based on 10 repeat-runs with different initialization of the network. This allows us to show the amount of variance in results we obtained

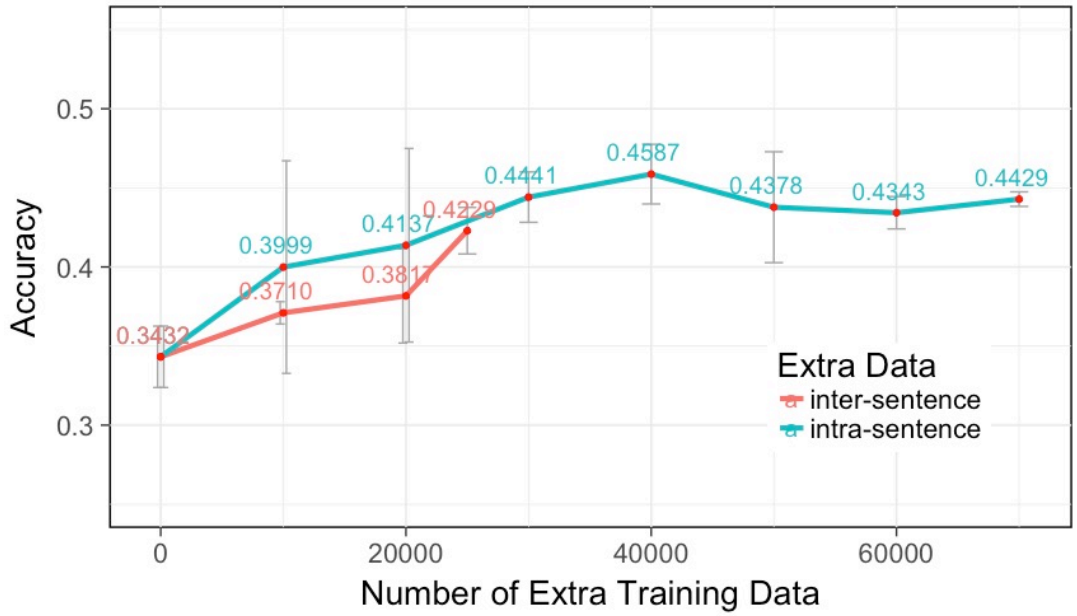

Figure 4.5: Average and variance of classification accuracy evaluated on the PDTB test set with different sample size.

in Figure 4.5. We found that results sometimes varied a lot between different runs, and would therefore like to encourage others in the field to also report variability due to initialization or other random factors. For instance, our best run achieved 49.84% accuracy on the PDTB test set trained with all additional instances, while mean performance for that setting is 45.50% accuracy. Variances were substantially smaller for the cross-validation setting, as the number of overall instances going into the evaluation is a lot larger in this setting, and hence yields more stable performance estimates.

4.5.2 Qualitative analysis

In order to illustrate what kinds of instances our method extracts, we show an instances below. The underlined DC is the explicit DC identified in the back-translated target text; the discourse relation is automatically classified based on the back-translation.

3. *[Justice demands it.]*$_{Arg1}$ ***but*** *[The minister refuses.]*$_{Arg2}$

— Comparison.Contrast

One strength of the proposed method is that it can mine and label discourse relations that are not commonly regarded as discourse relations and hence not annotated in PDTB. Below are some examples where the bold DC was identified in the (back-)translation:

4. *A conservative member was kicked out of his caucus for defending Nova Scotians.*

   — **because**, Contingency.Cause

5. *A failure to do so would affect our attitude to their eventual accession.*

   —***if**, Contingency.Condition*

6. *In January, Caronport 's mayor and volunteer fire chief, Royden Taylor, perished fighting a fire.*

   —***while**, Temporal.Synchronous*

7. *A full pension is payable after 40 years of residence in Canada.*

   —***if**, Contingency.Condition*

These extra samples are in fact an invaluable resource of discourse-informative patterns, which are not available to discourse relation parsers that are trained only on the PDTB dataset. These cases provide evidence that our proposed method can not only provide instances that are similar to implicit labelled instances, but detect additional patterns, as attempted in Mihăilă and Ananiadou (2014); Hidey and McKeown (2016) for causal relations, and generalize from the semantic content observed in such relations to actual implicit discourse relations.

For example, as reported in Section 4.5, numerous *Temporal* relations are mined from the parallel corpus. These include cases where the original text contained a verbal construction which expresses the temporal relation, which through back-translation gets expressed as a discourse relation, or where explicit relations include gerunds in the Arg2, e.g.

"any plan takes time to have the effect required"→ "before getting the effect required"

"how much longer do women have to wait for fairness?" → "before women have fairness."

"having gone over the estimates" → *"<u>after</u> going over the estimates."*

(source text followed by (back-)translation, where the explicitated DC is underlined).

In this work, we only extracted inter- and intra-sentential discourse relations, but the method can be in principle extended to other discourse relations that are not annotated in the PDTB, such as implicit relation between non-consecutive sentences. Discourse parsers that identify a larger range of relations are more useful in end applications. More importantly, identification of discourse-informative linguistic patterns by the proposed method opens the opportunity to mine extra samples under a monolingual setting and further improve classification performance.

4.5.3 Quantitative analysis

In order to get detailed insights on how much extra data is most beneficial to the task, we also trained our classifier with different numbers of additional extracted samples. Figure 4.5 compares the classification accuracy when training on incremental number of extra instances. We find that the performance increases with samples size, but plateaus after $40,000$ intra-sentential samples.

In fact, this sample size produces the highest averaged classification accuracy of 45.87%, which is even higher than our model which includes all extracted samples. A possible reason for not seeing further improvement by adding more intra-sentential examples is the difference in distribution and properties of these extra samples compared to the PDTB data. We also experimented with training on the parallel-text samples only (i.e., without any PDTB training samples), but the result was worse than using PDTB only. Adding more *inter-sentential* samples might further improve the performance, as these instances are closer to the PDTB data.

4.6 Methodological discussion

Our proposed method uses back-translated target discourse connectives to label implicit relations. The quality of the relation label is intrinsically subject to the translation policy of

the parallel corpora and also extrinsically subject to the accuracy of explicit DC classification by the end-to-end parser and the quality of the MT system. For example, a particularly high proportion of *Contingency.Condition* relations is found in the intra-sentential samples. Analyzing these samples, we found numerous instances where the word 'if' is wrongly identified as a DC (e.g. *He asked if it was correct.*). It is not surprising to have noisy samples extracted because limited screening strategy is applied in the current method.

As a reference for the quality of the relation label produced, we analysed the intra-sentential relations in the parallel corpus that are explicit on the source side and also in the back-translation. We found that 68% of the originally explicit DCs are (back-)translated to the same explicit DCs and 75% to DCs of the same level-2 sense, according to automatic explicit DC classification of the end-to-end parser.

4.7 Summary

In this chapter, we show that explicitation during human translation can provide a valuable signal for expanding datasets for implicit discourse relations. As the expansion of training instances is orthogonal to the mechanism of implicit discourse relation classification, this method can be applied to improve any methods of implicit discourse relation classification.

Compared to previous methods, the proposed model is much simpler and more practical in reality. Our best run in fact even reaches 49.84% accuracy on the PDTB test set. In Chapter 3, I point out that small evaluation set is too risky to draw conclusions in this task, we still get significant improvements on the cross validation setting which tells that the method is effective and additional data we got indeed helped.

There is plenty of room for further improvement by controlling the sample quality, such as selection based on explicit discourse connective identification confidence, restraining the discourse relation structure, identifying Arg1 and Arg2 such that approaches which use two separate representations for arguments instead of a single concatenated vector become possible, reducing language-specific bias by mining from parallel corpora of other language pairs, and fine-tuning the MT system for discourse connective translation. We leave the exploration of these areas to the next chapter. We also experiment after excluding samples with

ambiguous connectives in the (back-)translation, such as "and", as the discourse connectives classification is likely to be nosier. However, the resulting implicit relation classification is slightly degraded.

To alleviate the above problems, in the next chapter, I expand the language pairs to German, French and Czech and also re-paragraph the texts to make the arguments coherent and in topic.

Chapter 5

Multilingual explicitation for implicit discourse relation classification

5.1 Introduction

In the last chapter I introduce a back-translation method, which exploits the fact that human translators sometimes insert a connective in their translation even when a relation is implicit in the original text. Using a back translation method, I showe that such instances can be used for acquiring additional automatically annotated texts.

However, in the last chapter, only a single target language (French) is used and we have no control over the quality of the labels extracted from back-translated connectives. In this chapter, I therefore systematically compare the contribution of three target translation languages from different language families: French (a Romance language), German (from the Germanic language family) and Czech (a Slavic language). As all three of these languages are part of the EuroParl corpus (Koehn, 2005), this also allows us to directly test whether higher quality can be achieved by using those instances that were consistently explicitated in several languages. We use cross-lingual explicitation to acquire more reliable implicit discourse relation instances with separate arguments that are from adjacent sentences in a document,

and conduct experiments on PDTB benchmark with multiple conventional settings including cross validation. The experimental results show that the performance can be improved significantly with the additional training data, compared with the baseline systems.

5.2 Methodology

Our goal here aims at sentence pairs in cross-lingual corpora where connectives have been inserted by human translators during translation from English to several other languages. After automatically back-translating from other languages to English, explicit relations can be easily identified by a discourse parser and then original English sentences would be labeled accordingly.

We follow the pipeline proposed in Chapter 4, as illustrated in Figure 5.1, with the following differences:

- The model in Chapter 4 suffers from the fact that typical sentence-aligned corpora may have some sentences removed and make the sentences no longer coherent to get inter-sentential discourse relation instances. Here we filter and re-paragraph the line-aligned corpus to parallel document-aligned files, which makes it possible to obtain in-topic inter-sentential instances. After preprocessing, we got 532,542 parallel sentence pairs in 6,105 documents.
- In Chapter 4, I point out that having correct translation of explicit discourse connective is more important than having the correct translation of the whole sentence. In this chapter I use a statistical machine translation system instead of a neural one for more stable translations of discourse connectives.
- Instead of a single language pair, we use three language pairs and majority votes between them to get annotated implicit discourse relation instances with high confidence.

Figure 6.1 illustrates the pipeline of our approach. It consists of a few steps including preprocessing, back-translating, discourse parsing and majority voting. For each document, we back-translate its German, French and Czech translation back to English with the MT

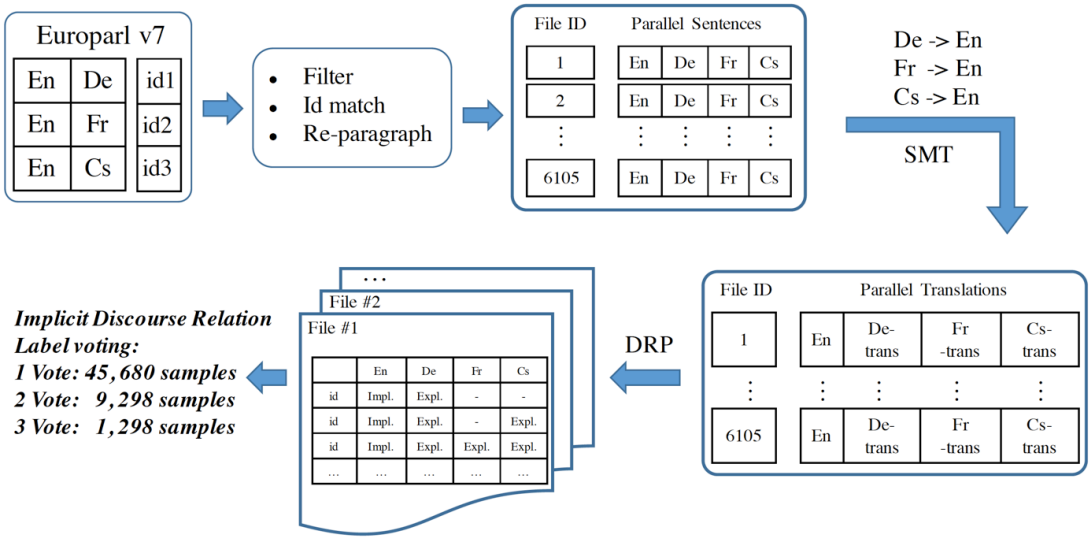

Figure 5.1: The pipeline of proposed method. "SMT" and "DRP" denote statistical machine translation and discourse relation parser respectively.

system and parse them with discourse parser. In this way, we can easily identify those instances that are originally implicit but explicit in the human translations to German, French or Czech. With majority vote by the explicit examples in those three languages, the original English instance can be labeled, and labelling confidence can be estimated.

5.2.1 Preprocessing

We use European Parliament Proceedings Parallel Corpus (Europarl[1]) (Koehn, 2005) and choose English-French, English-German and English-Czech pairs as our parallel corpora. Each source-target pair consists of source and target sentences along with a sentence ID with which we can easily identify the location of the sentence in its paragraph. In order to get document-aligned parallel sentences between all these four languages, we do preprocessing steps as follows:

- Filtering: remove those sentences that don't have all the three translations in French, German or Czech. In this way, each input sentence pair would have the same votes in the later steps.

- ID matching: re-group each sentence into its origin document by the sentence IDs.

[1]Data is downloaded from `http://opus.nlpl.eu/Europarl.php`

- Re-paragraph: order the sentences in each documents by the ID and re-paragraph them. This is to make the arguments of the extracted implicit discourse instance consecutive and are in the same context.

5.2.2 Machine translation

We train three MT systems to back-translate French, German and Czech to English. To have word alignments, better and stable back-translations, we employ a statistical machine translation system MOSES[2] (Koehn et al., 2007), trained on the same parallel corpora. Source and target sentences are first tokenized, true-cased and then fed into the system for training. In our case, the translation target texts are identical with the training set of the translation systems; this would not be a problem because our only objective in the translation is to back-translate connectives in the translation into English. On the training set, the translation system achieves BLEU scores of 66.20 (French), 65.30 (German) and 69.05 (Czech), while on **news2015 and newsdiscusstest2015** it achieves BLEU scores of 23.92[3], 17,93 and 18.82 respectively. The reason why they performed much worse on the test set than the current state-of-the-art (namely 35.00), is that they are only trained on a small part of the normal training set in MT practice. In this case, it doesn't matter too much because we only need to make sure that the implicit connectives in other languages can be back-translated to English correctly.

5.2.3 Discourse parser

We employ the PDTB-style parser proposed in Lin et al. (2014), which achieved about 96% accuracy on explicit connective identification, to pick up those explicit examples in back-translations in each document. Following the definitions of discourse relations in the PDTB that the arguments of the implicit discourse relations should be adjacent sentences but not for the explicit relations, we screen out all those explicit samples from the outputs of the parser that don't have consecutive arguments.

[2] `http://www.statmt.org/moses/`

[3] Case sensitive BLEU implemented in *meteval-v13a.pl*.

5.2.4 Majority vote

After parsing the back-translations of French, German and Czech, we can compare whether they contain explicit relations which connect the same relational arguments. The analysis of this subset then allows us to identify those instances that can be labeled with high confidence, i.e. where back-translations from all three languages allow us to infer the same coherence label. Note that it is not necessarily the case that all back-translations contain an explicitation for the same instance (for instance, the French translator may have explicitated a relation, while the German and the Czech translators didn't do so), or that they propose *the same* coherence label: the human translation can introduce "noise" in the sense of the human translators inferring different coherence relations, the machine translation model can introduce errors in back-translation, and the discourse parser can mislabel ambiguous explicit connectives. When we use back-translations of several languages, the idea is that we can eliminate much of this noise by selecting only those instances where all back-translations agree with one another, or the ones where at least two back-translations allow us to infer identical labels.

Figure 5.2 illustrates the number of automatically labeled implicit discourse relation examples together with the information of how many of the instances that just one, two or all three back-translations provided the same labels.

In the One Vote agreement, every explicit relation has been accepted and the original implicit English sentences have been annotated correspondingly. Likewise, Two Votes agreement needs at least two out of three languages to have the same explicit relation label after back-translation; agreement between all three back-translations is denoted as Three Votes.

5.3 Experiments

5.3.1 Data

Europarl Corpora: The parallel corpora used here are from Europarl (Koehn, 2005), it contains about 2.05M English-French, 1.96M English-German and 0.65M English-Czech pairs. After preprocessing, we got about 0.53M parallel sentence pairs in all these four languages.

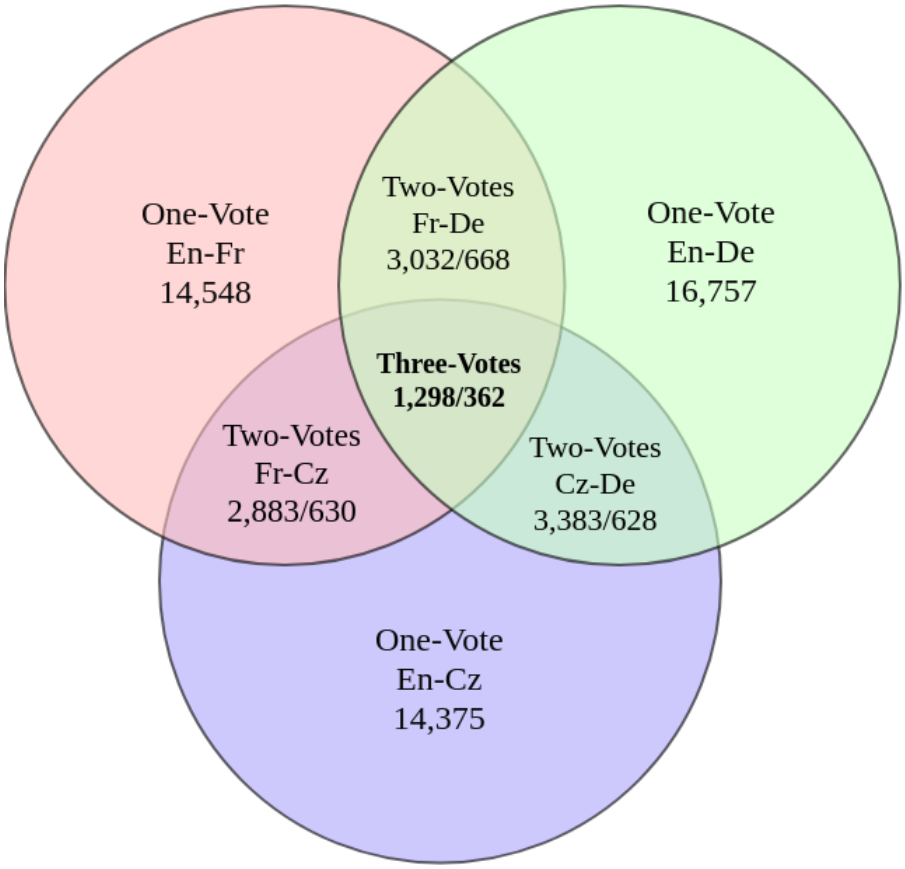

Figure 5.2: Numbers of implicit discourse relation instances from different agreements of explicit instances in three back-translations. En-Fr denotes instances that are implicit in English but explicit in back-translation of French, same for En-De and En-Cz. The overlap means they share the same relational arguments. The numbers under "Two-Votes" and "Three-Votes" are the numbers of discourse relation agreement / disagreement between explicits in back-translations of two or three languages.

The Penn Discourse Treebank (PDTB): PDTB (Prasad et al., 2008) is the largest available manually annotated corpus of discourse relations from Wall Street Journal, as introduced in Chapter 2. In this chapter, I follow the previous conventional settings and focus on the second-level 11-ways classification (Lin et al., 2009; Ji and Eisenstein, 2015; Rutherford et al., 2017b; Shi et al., 2017), after removing the relations with few instances.

5.3.2 Implicit discourse relation classification

To evaluate whether the extracted data is helpful to this task, we use a simple and effective bidirectional Long Short-Term Memory (LSTM) (Hochreiter and Schmidhuber, 1997)

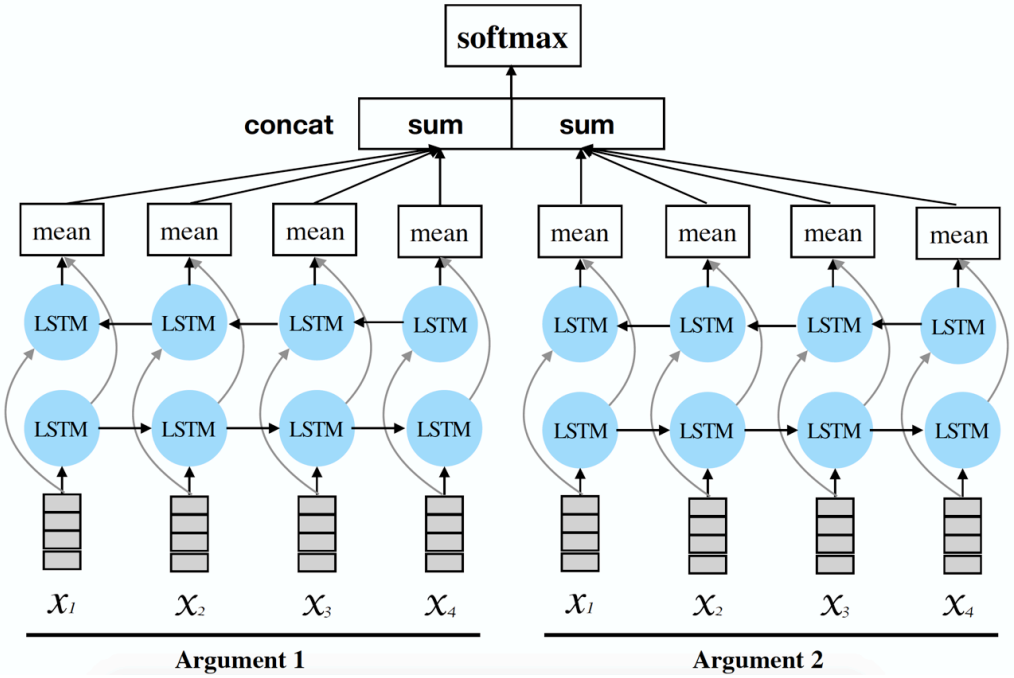

Figure 5.3: Bi-LSTM network for implicit discoure relation classification.

network.

A LSTM recurrent neural network processes a variable-length sequence $x = (x_1, x_2, ..., x_n)$. At time step t, the state of memory cell c_t and hidden h_t are calculated with the Equations 5.1:

$$\begin{bmatrix} i_t \\ f_t \\ o_t \\ \hat{c}_t \end{bmatrix} = \begin{bmatrix} \sigma \\ \sigma \\ \sigma \\ \tanh \end{bmatrix} W \cdot [h_{t-1}, x_t] \tag{5.1}$$

$$c_t = f_t \odot c_{t-1} + i_t \odot \hat{c}_t$$

$$h_t = o_t \odot \tanh(c_t)$$

After being mapped to vectors, words are fed into the network sequentially. Hidden states of LSTM cell from different directions are averaged. The representations of two arguments from two separate bi-LSTMs are concatenated before being fed into a softmax layer for prediction. The architecture is illustrated in Figure 5.3.

Implementation: The model is implemented in Pytorch[4]. All the parameters are initialized

[4] `https://pytorch.org/`

uniformly at random. We employ cross-entropy as our cost function, Adagrad with learning rate of 0.01 as the optimization algorithm and set the dropout layers after embedding and output layer with drop rates of 0.5 and 0.2 respectively. The word vectors are pre-trained word embeddings from Word2Vec[5].

Settings: We follow the previous works and evaluate our data on second-level 11-ways classification on PDTB with 3 settings: Lin et al. (2009) (denotes as PDTB-Lin) uses sections 2-21, 22 and 23 as train, dev and test set; Ji and Eisenstein (2015) (denotes as PDTB-Ji) uses sections 2-20, 0-1 and 21-22 as train, dev and test set; Moreover, I also use 10-folds cross validation among sections 0-23 (Shi and Demberg, 2017). For each experiment, the additional data is only added into the training set.

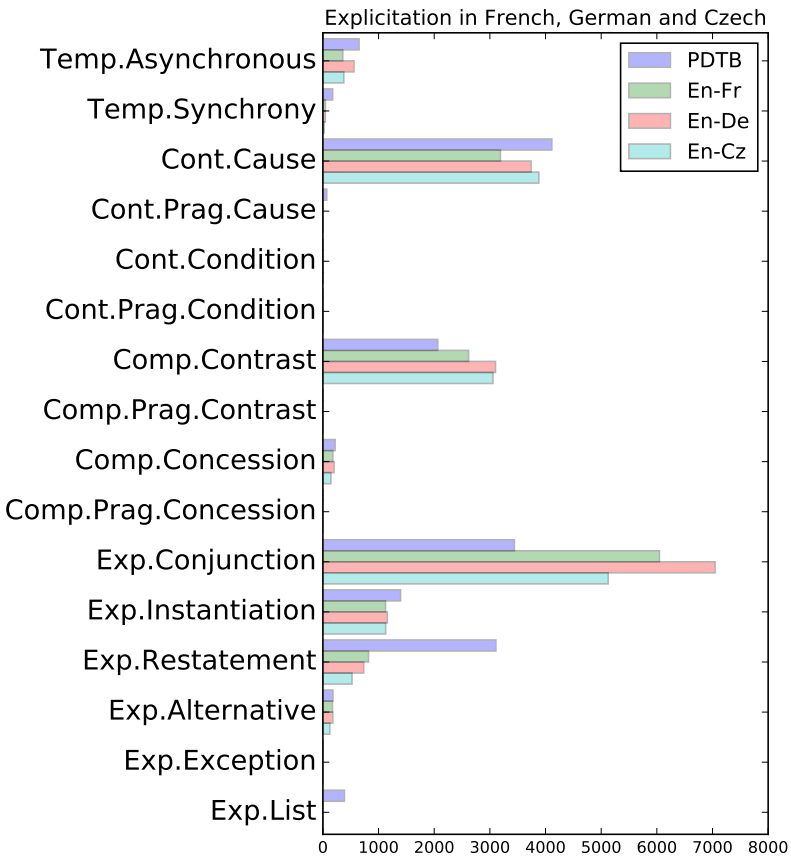

Figure 5.4: Distributions of PDTB and the extracted data among each discourse relation.

[5]https://code.google.com/archive/p/word2vec/

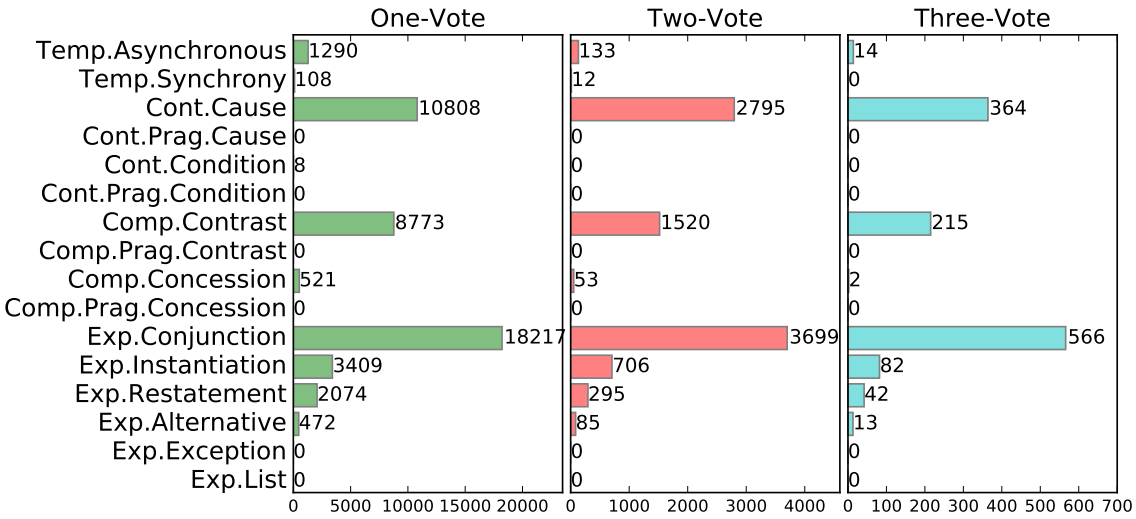

Figure 5.5: Distributions of discourse relations with different agreements.

		PDTB-Lin	PDTB-Ji	Cross Validation	size of extra data
Majority Class		26.11	26.18	25.59	-
Rutherford et al. (2017b)		38.38	-	-	-
Shi et al. (2017)		**45.50**	-	37.84	102,314
PDTB only		37.95(0.59)	40.57(0.67)	37.82(0.14)	-
PDTB +	En-Fr	38.96(0.69)	40.14(0.78)	38.32(0.62)	14,548
	En-De	39.65(0.95)	39.96(0.44)	37.97(0.46)	16,757
	En-Cz	37.90(1.27)	40.59(0.51)	37.42(0.50)	14,375
	All	37.73(0.74)	40.41(0.65)	37.16(0.64)	45,680
PDTB + 2-votes		**40.34**(0.75)	**41.95**(0.97)	**38.98**(0.14)	9,298
PDTB + 3-votes		39.88(0.79)	41.19(0.63)	38.33(0.50)	1,298

Table 5.1: Performances with different sets of additional data. Average accuracy of 10 runs (5 for cross validations) are shown here with standard deviation in the brackets. Numbers in bold are significantly ($p<0.05$) better than the *PDTB only* baseline with unpaired t-test.

5.4 Results and analysis

5.4.1 Distribution of new instances

Figure 5.4 shows the distributions of expert-annotated PDTB implicit relations and the implicit discourse examples extracted from the French, German and Czech back-translations. Overall, there is no strong bias – all relations seem to be represented similarly well, in line

with their general frequency of occurrence. One interesting exception is the higher number of *Expansion.Conjunction* relation from the German translations. The over-representation of *Expansion.Conjunction* relation in German indicates that German translators tend to use more explicit cues to mark these relations. This is an independently discovered well-known finding from the literature (Kunz and Lapshinova-Koltunski, 2015), which observed that German tends to mark conjunction relations with discourse cues, while English tends to use coreference instead. We also find that *Expansion.Restatement* relations are under-represented in our back-translation method, indicating that these relations are explicitated particularly rarely in translation. We also find that we can identify more *Contingency.Cause* and *Comparison.Contrast* relations from the German and Czech back-translations compared to the French ones. This provides us with an interesting lead for future work, to investigate whether French tends to explicitate these relations less, expressing them implicitly like in the English original, or whether French connectives for causal and contrastive relations are more ambiguous, causing problems in the back-translations.

Figure 5.5 shows that the filtering by majority votes (including only two cases where at least two back-translations agree with one another vs. where all three agree) does again not change the distribution of extracted relations.

In summary, we can conclude that the choice of translation language *can* matter: depending on what types of relations are most important to acquire more data for the target task at hand, a language that tends to explicitate that relation frequently can be particularly suitable. On the other hand, if no strong such preferences on labelling specific relations exist, we can see that the choice of translation language only has a minor effect on the overall distribution of additional implicit discourse relation labels.

5.4.2 Quantitative results

Table 5.1 shows that best results are achieved by adding only those samples for which two back-translations agree with one another. This may represent the best trade-off between reliability of the label and the amount of additional data. The setting where the data from all languages is added performs badly despite the large number of samples, because this method contains different labels for the same argument pairs, for all those instances where

the back-translations don't yield the same label, thus introducing noise into the system. The size of the extra data used in Chapter 4 is about 10 times larger than our 2-votes data. The selection of instances differs in their paper from ours, in that they only use French, and in that they, unlike this chapter, focus on intra-sentential samples. The model using the few reliable samples extracted from the back-translations of the three languages here significantly outperforms the results in Chapter 4 in the cross-validation setting. On the PDTB-Lin test set, we don't match performance, but note that this test set is based only on 800 instances, as opposed to the 16k instances in the cross-validation evaluation.

5.4.3 Qualitative analysis

Finally, we want to provide insight into what kind of instances the system extracts, and why back-translation labels sometimes disagree. We have identified four major cases based on a manual analysis of 100 randomly sampled instances.

Case 1: Sometimes, back-translations from several languages may yield the same connective because the original English sentence actually was not really unmarked, but rather contained an expression which could not be automatically recognized as a discourse relation marker by the automatic discourse parser[6]. This can actually help us to identify new alternative lexicalization for discourse relations, and thus represents a promising technique for improving discourse relation classification also on texts for which no translations are available.

Original English: I presided over a region crossed by heavy traffic from all over Europe, with significant accidents which gave rise to legal actions. ***What is more,*** In 2002, two Member States of the European Union appealed to the European Court of Justice to repeal Directive 2002/15/EC because it included self-employed drivers ; the Court rejected their appeal on the grounds of road safety.

French back-translation: I presided over a region crossed by heavy traffic from the whole of Europe, with significant accidents which gave rise to legal actions, **moreover** in 2002 , two Member States have appeal on the European Court of Justice, which has condemned

[6] In the following examples, the original English sentence is shown is followed by the back-translations from French, German and Czech along with the connectives and senses.

the rejection of the grounds of road safety.

(Expansion.Conjunction)

German back-translation: I presided over a region crossed by heavy traffic from across Europe, with significant accidents which, **moreover** in 2002, two Member States of the European Union appealed to the European Court of Justice to repeal Directive 2002/15/EC , because it included self-employed drivers ; the Court quashed for reasons of road safety.

(Expansion.Conjunction)

Czech back-translation: I was in the region with very heavy traffic from all over Europe, with significant accidents which gave rise to legal actions **therefore** after all, in 2002, two Member States of the European Union appealed to the European Court of Justice to repeal Directive 2002/15/EC that also applies to self-employed drivers; the Court rejected their appeal on the grounds of road safety.

(Contingency.Cause)

The expression *what is more* is not part of the set of connectives labeled in PDTB and hence was not identified by the discourse parser. Our method is successful because such cues can be automatically identified from the consistent back-translations into two languages. (The case in Czech is more complex because the back-translation contains two signals, *therefore* and *after all*, see case 4.)

We also found some similar expressions in this case like:

"in reality" ("implicit", original English) → "in fact" (explicit, back-translation);

"for that reason" → "therefore";

"this is why" → "therefore";

"be that as it may" → "however / nevertheless";

"for another" → "furthermore / on the other hand";

"in spite of that" → "however / nevertheless" and so on.

Case 2: Majority votes help to reduce noise related to errors introduced by the automatic pipeline, such as argument or connective misidentification: in the below example, *also* in the French translation is actually the translation of *along with*.

Original English: on behalf of the PPE-DE Group. (DE) Madam President, Commissioner, ladies and gentlemen, the public should be able to benefit in two ways from the potential for greater road safety. ***For this reason***, along with the report we are discussing today, I call for more research into the safety benefits of driver-assistance systems.

French back-translation: (DE) Madam President, Commissioner, ladies and gentlemen, citizens should be able to benefit in two ways of the possibility of improving road safety. **also** when we are discussing this report today, I appeal to the intensification of research at the level of the benefits of driver-assistance systems in terms of security, as well as the transmission of information about them.

(Expansion.Conjunction)

German back-translation: (DE) Madam President, Commissioner, ladies and gentlemen, road safety potentials should citizens in the dual sense **therefore** I urge, together with the report under discussion today, the prevention and education about the safety benefits of driver-assistance systems.

(Contingency.Cause)

Czech back-translation: (DE) Madam President, Commissioner, ladies and gentlemen, the public would be the potential for greater road safety should have a two-fold benefit, **therefore** I call, in addition to the report, which we are debating today , for more research and education in the safety benefits of driver-assistance systems.

(Contingency.Cause)

Case 3: Discrepancies between connectives in back-translations can also be due to differences in how translators interpreted the original text. Here are cases of genuine ambiguities in the implicit discourse relation.

Original English: with regard, once again, to European Union law, we are dealing in this case with the domestic legal system of the Member States. ***That being said***, I cannot answer for the Council of Europe or for the European Court of Human Rights, which have issued a decision that I understand may raise some issues for Parliament.

French back-translation: with regard, once again, the right of the European Union, we are here in the domestic legal system of the Member States. **however,** I cannot respond to the

place of the Council of Europe or for the European Court of Human Rights, which have issued a decision that I understand may raise questions in this House.

(Comparison.Contrast)

German back-translation: once again on the right of the European Union, we have it in this case with the national legal systems of the Member States. **therefore,** I cannot, for the Council of Europe and the European Court of Human Rights, which have issued a decision, which I can understand, in Parliament raises some issues.

(Contingency.Cause)

Czech back-translation: I repeat that, when it comes to the European Union, in this case we are dealing with the domestic legal system of the Member States. **in addition,** I cannot answer for the Council of Europe or for the European Court of Human Rights , which has issued a decision that I understand may cause in Parliament some doubts.

(Expansion.Conjunction)

Case 4: Implicit relations can co-occur with marked discourse relations (Rohde et al., 2015), and multiple translations help discover these instances, for example:

Original English: We all understand that nobody can return Russia to the path of freedom and democracy, ***(implicit: but) what is more***, the situation in our country is not as straightforward as it might appear to the superficial observer.

French back-translation: we all understand that nobody can return Russia on the path of freedom and democracy but Russia itself, its citizens and its civil society **but** there is more, the situation in our country is not as simple as it might appear to be a superficial observer.

(Comparison.Contrast)

German back-translation: we are all aware that nobody Russia back on the path of freedom and democracy, as the country itself, its people and its civil society **but** the situation in our country is not as straightforward as it might appear to the superficial observer.

(Comparison.Contrast)

Czech back-translation: we all know that Russia cannot return to the path of freedom and democracy there, but Russia itself, its people and civil society. **in addition** the situation in

our country is not as straightforward as it might appear to the superficial observer.

(Expansion.Conjunction)

5.5 Summary

In this chapter, I compare the explicitations obtained from translations into three different languages, and find that instances where at least two back-translations agree yield the best quality, significantly outperforming a version of the model that does not use additional data, or uses data from just one language.

I also found that specific properties of the translation language affect the distribution of the additionally acquired data across coherence relations: German, for instance, is known to mark conjunction relations using discourse cues more frequently, while English and other languages tend to express these relations rather through lexical cohesion or pronouns. This was reflected in our experiments: we found a larger proportion of explicitations for conjunction relations in German than the other translation languages.

Finally, the qualitative analysis shows that the strength of the method partially stems from being able to learn additional discourse relation signals because these are typically translated consistently. The method thus shows promise for the identification of discourse markers and alternative lexicalizations, which can subsequently be exploited also for discourse relation classification in the absence of translation data. Our analysis also shows that our method is useful for identifying cases where multiple relations holding between two arguments.

The idea of using the explicitation process in human translating not only can be applied with multi-lingual data, but also it can be used in monolingual data and making better use of the implicit connective annotations in the PDTB with neural networks (Qin et al. (2017) shows that the implicit connective annotation also helps neural networks in learning better sentence representations). Moreover, with the limited number of data, learning the surface cues is obviously not adequate for tasks that need deeper encoding and interpretation. Having better understanding of how arguments relate to one another and having better semantic representation are crucial. In the next chapter, I will introduce a sequence-to-sequence

neural network based method, which forces the internal representation to more completely encode the semantics of the relational arguments, and makes a more fine-grained classification than is necessary for the overall task.

Chapter 6

Learning to explicitate connectives with a Seq2Seq network

In Chapter 3 we conclude that having better understanding of how two relation arguments are related to one another and better semantic sentence representations are of crucial importance to the task. The implicit connective annotations in the PDTB give us the chance to transfer the idea of using the explicitation process in human translating to monolingual data. In this chapter, we propose a sequence-to-sequence network model to mimic the process of explicitation and the evaluate the approach on different settings and data splits.

6.1 Introduction

The Penn Discourse Tree Bank (Prasad et al., 2008, PDTB) provides lexically-grounded annotations of discourse relations and their two discourse relational arguments (i.e., two text spans).

When annotating implicit relations in the PDTB, annotators were asked to first insert a connective which expresses the relation, and then annotate the relation label. This procedure was introduced to achieve higher inter-annotator agreement for implicit relations between human annotators. In the approach taken in this chapter, our model mimics this procedure

by being trained to explicitate the discourse relation, i.e. to insert a connective as a secondary task.

The key in implicit discourse relation classification lies in extracting relevant information for the relation label from (the combination of) the discourse relational arguments. Informative signals can consist of surface cues, as well as the semantics of the relational arguments. Statistical approaches have typically relied on linguistically informed features which capture both of these aspects, like temporal markers, polarity tags, Levin verb classes and sentiment lexicons, as well as the Cartesian products of the word tokens in the two arguments (Lin et al., 2009). More recent efforts use distributed representations with neural network architectures (Qin et al., 2016a).

The main question in designing neural networks for discourse relation classification is how to get the neural networks to effectively encode the discourse relational arguments such that all of the aspects relevant to the classification of the relation are represented, in particular in the face of very limited amounts of annotated training data, see e.g. Rutherford et al. (2017b). The crucial intuition in the present chapter is to make use of the annotated implicit connectives in the PDTB: in addition to the typical relation label classification task, we also train the model to encode and decode the discourse relational arguments, and at the same time predict the implicit connective. This novel secondary task forces the internal representation to more completely encode the semantics of the relational arguments (in order to allow the model to decode later), and to make a more fine-grained classification (predicting the implicit connective) than is necessary for the overall task. This more fine-grained task thus aims to force the model to represent the discourse relational arguments in a way that allows the model to also predict a suitable connective.

Our overall discourse relation classifier combines representations from the relational arguments as well as the hidden representations generated as part of the encoder-decoder architecture to predict relation labels. What's more, with an explicit memory network, the network also has access to history representations and acquire more explicit context knowledge. We show that our method outperforms previous approaches on the 11-way classification on the PDTB 2.0 benchmark.

6.2 System overview

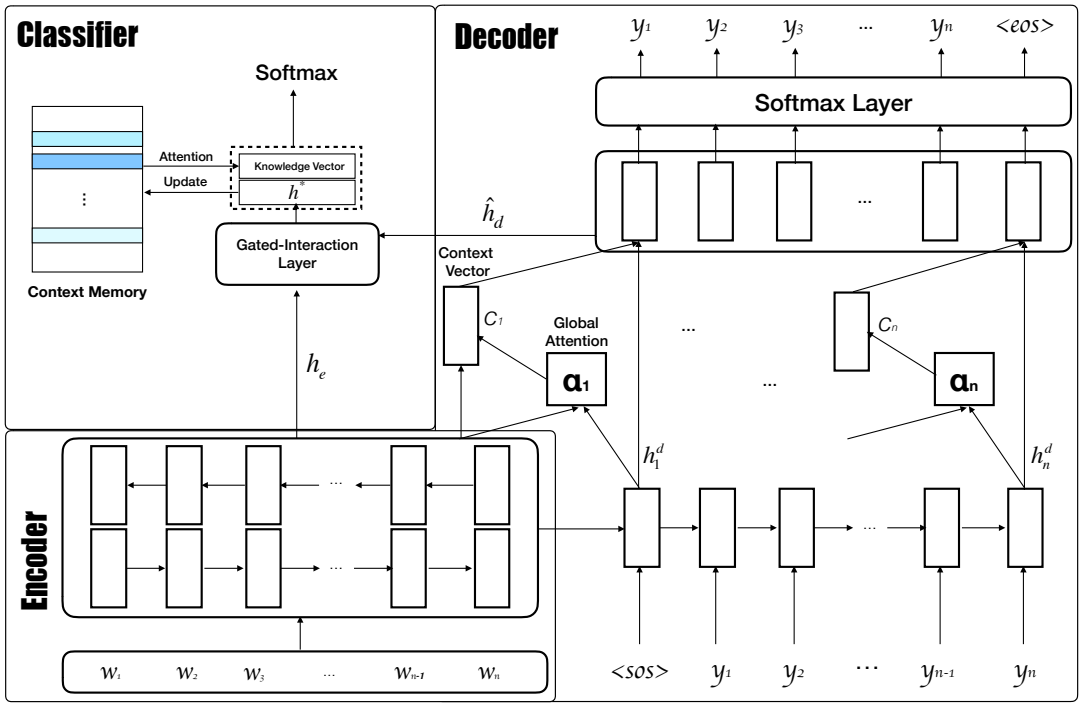

Figure 6.1: The Architecture of Proposed Model.

Our model is based on the sequence-to-sequence model used for machine translation (Luong et al., 2015), an adaptation of an LSTM (Hochreiter and Schmidhuber, 1997) that encodes a variable length input as a fix-length vector, then decodes it into a variable length of outputs. As illustrated in Figure 6.1, our model consists of three components: Encoder, Decoder and Discourse Relation Classifier. We here use different LSTMs for the encoding and decoding tasks to help keep the independence between those two parts.

The task of implicit discourse relation recognition is to recognize the senses of the implicit relations, given the two arguments. For each discourse relation instance, The Penn Discourse Tree Bank (PDTB) provides two arguments (Arg_1, Arg_2) along with the discourse relation (*Rel*) and manually inserted implicit discourse connective ($Conn_i$). Here is an implicit example from section 0 in PDTB:

3. **Arg$_1$**: This is an old story.

   Arg$_2$: We're talking about years ago before anyone heard of asbestos having any questionable properties.

Conn$_i$: in fact

Rel: Expansion.Restatement

During training, the input and target sentences for sequence-to-sequence neural network are $[Arg_1; Arg_2]$ and $[Arg_1; Conn_i; Arg_2]$ respectively, where ";" denotes concatenation.

6.3 Model components

6.3.1 Encoder

Given a sequence of words, an encoder computes a joint representation of the whole sequence.

After mapping tokens to Word2Vec embedding vectors (Mikolov et al., 2013), a LSTM recurrent neural network processes a variable-length sequence $x = (x_1, x_2, ..., x_n)$ by incrementally adding new contents into a single memory cell, with gates controlling the content to which contents should be memorized, erased or inputed. At time step t, the state of memory cell c_t and hidden h_t are calculated with the Equations 6.1:

$$\begin{bmatrix} i_t \\ f_t \\ o_t \\ \hat{c}_t \end{bmatrix} = \begin{bmatrix} \sigma \\ \sigma \\ \sigma \\ \tanh \end{bmatrix} W \cdot [h_{t-1}, x_t]$$
$$c_t = f_t \odot c_{t-1} + i_t \odot \hat{c}_t \tag{6.1}$$
$$h_t = o_t \odot \tanh(c_t)$$

where x_t is the input at time step t, i, f and o are the input, forget and output gate activation respectively. $\hat{c}_t$ denotes the current cell state, σ is the logistic sigmoid function and $\odot$ denotes element-wise multiplication. The LSTM separates the memory c from the hidden state h, which allows for more flexibility incombining new inputs and previous context.

For the sequence modeling tasks, it is beneficial to have access to the past context as well as the future context. Therefore, we chose a bidirectional LSTM as the encoder and the output

of the word at time-step t is shown in the Equation 6.2. Here, element-wise sum is used to combine the forward and backward pass outputs.

$$h_t = \left[\overrightarrow{h_t} \oplus \overleftarrow{h_t}\right] \tag{6.2}$$

Thus we get the output of encoder:

$$h_e = [h_1^e, h_2^e, ..., h_n^e] \tag{6.3}$$

6.3.2 Decoder

With the representation from the encoder, the decoder tries to map it back to the targets space and predicts the next words.

Here we used a separate LSTM recurrent network to predict the target words. During training, target words are fed into the LSTM incrementally and we get the outputs from decoder LSTM:

$$h_d = \left[h_1^d, h_2^d, ..., h_n^d\right] \tag{6.4}$$

Global attention

In each time-step in decoding, rather than choosing the hidden state of the last time-step, it's better to consider all the hidden states of the encoder to give the decoder a full view of the source context. So we adopt the global attention mechanism proposed in Luong et al. (2015). For time step t in decoding, context vector c_t is the weighted average of h_e, the weights for each time-step are calculated with h_t^d and h_e as illustrated below:

$$\alpha_t = \frac{\exp({h_t^d}^\top \mathbf{W}_\alpha h_e)}{\sum_{t=1}^{n} \exp({h_t^d}^\top \mathbf{W}_\alpha h_e)} \tag{6.5}$$

$$c_t = \alpha h_e \tag{6.6}$$

Word prediction

Context vector c_t captures the relevant source side information to help predict the current target word y_t. We employ a concatenate layer with activation function tanh to combine context vector c_t and hidden state of decoder h_t^d at time-step t as follows:

$$\hat{h_t^d} = \tanh(\mathbf{W_c}\left[c_t; h_t^d\right]) \tag{6.7}$$

Then the predictive vector is fed into the softmax layer to get the predicted distribution $\hat{p}(y_t|s)$ of the current target word.

$$\begin{aligned} \hat{p}(y_t|s) &= softmax(\mathbf{W}_s\hat{h_d} + \mathbf{b}_s) \\ \hat{y_t} &= \arg\max_y \hat{p}(y_t|s) \end{aligned} \tag{6.8}$$

After decoding, we obtain the predictive vectors for the whole target sequence

$$\hat{h_d} = \left[h_1^d, h_2^d, ..., h_n^d\right] \tag{6.9}$$

Ideally, it contains the information of exposed implicit connectives.

Gated interaction

In order to predict the coherent discourse relation of the input sequence, we take both the $h_{encoder}$ and the predictive word vectors h_d into account. K-max pooling can "draw together" features that are most discriminative and among many positions apart in the sentences, especially on both the two relational arguments in our task here; this method has been proved to be effective in choosing active features in sentence modeling (Kalchbrenner et al., 2014). We employ an average k-max pooling layer which takes average of the top k-max values among the whole time-steps as in Equation 6.10 and 6.11:

$$\bar{h}_e = \frac{1}{k}\sum_{i=1}^{k} topk(h_e) \tag{6.10}$$

$$\bar{h}_d = \frac{1}{k}\sum_{i=1}^{k} topk(\hat{h^d}) \tag{6.11}$$

$\bar{h}_e$ and $\bar{h}_d$ are then combined using a linear layer (Lan et al., 2017). As illustrated in Equation 6.12, the linear layer acts as a gate to determine how much information from the sequence-to-sequence network should be mixed into the original sentence's representations from the encoder. Compared with bilinear layer, it also has less parameters and allows us to use high dimensional word vectors.

$$h^* = \bar{h}_e \oplus \sigma(\mathbf{W}_i \bar{h}_d + \mathbf{b}_i) \tag{6.12}$$

6.3.3 Explicit context knowledge

To further capture common knowledge in contexts, we here employ a memory network proposed in Liu et al. (2018), to get explicit context representations of contexts training examples. We use a memory matrix $M \in R^{K\times N}$, where K, N denote hidden size and number of training instances respectively. During training, the memory matrix remembers the information of training examples and then retrieves them when predicting labels.

Given a representation h^* from the interaction layer, we generate a **knowledge vector** by weighted memory reading:

$$k = Msoftmax(M^T h^*) \tag{6.13}$$

We here use dot product attention, which is faster and space-efficient than additive attention, to calculate the scores for each training instances. The scores are normalized with a softmax layer and the final knowledge vector is a weighted sum of the columns in memory matrix M. Afterwards, the model predicts the discourse relation using a softmax layer.

$$\begin{aligned} \hat{p}(r|s) &= softmax(\mathbf{W}_r[k; h^*] + \mathbf{b}_r) \\ \hat{r} &= \arg\max_y \hat{p}(r|s) \end{aligned} \tag{6.14}$$

6.3.4 Multi-objectives

In our model, the decoder and the discourse relation classifier have different objectives. For the decoder, the objective consists of predicting the target word at each time-step. The loss function is calculated with masked cross entropy with `L2` regularization, as follows:

$$Loss_{de} = -\frac{1}{n}\sum_{t=1}^{n} y_t \log(\hat{p_y}) + \frac{\lambda}{2} \parallel \theta_{de} \parallel_2^2 \tag{6.15}$$

where y_t is one-hot represented ground truth of target words, $\hat{p_y}$ is the estimated probabilities for each words in vocabulary by softmax layer, n denotes the length of target sentence. λ is a hyper-parameter of $L2$ regularization and θ is the parameter set.

The objective of the discourse relation classifier consists of predicting the discourse relations. A reasonable training objective for multiple classes is the categorical cross-entropy loss. The loss is formulated as:

$$Loss_{cl} = -\frac{1}{m}\sum_{i=1}^{m} r_i \log(\hat{p_r}) + \frac{\lambda}{2} \parallel \theta_{cl} \parallel_2^2 \tag{6.16}$$

where r_i is one-hot represented ground truth of discourse relation labels, $\hat{p_r}$ denotes the predicted probabilities for each relation class by softmax layer, m is the number of target classes. Just like above, λ is a hyper-parameter of $L2$ regularization.

For the overall loss of the whole model, we set another hyper-parameter w to give these two objective functions different weights. Larger w means that more importance is placed on the decoder task.

$$Loss = w \cdot Loss_{de} + (1 - w) \cdot Loss_{cl} \tag{6.17}$$

6.4 Experiments and results

6.4.1 Experimental setup

We evaluate our model on the PDTB. While early work only evaluate classification performance between the four main PDTB relation classes, more recent work including the CoNLL

Settings	Train	Dev	Test
PDTB-Lin	13351	515	766
PDTB-Ji	12826	1165	1039
Cross valid. per fold avg.	12085	1486	1486[1]

Table 6.1: Numbers of train, development and test set on different settings for 11-way classification task. Instances annotated with two labels are double-counted and some relations with few instances have been removed.

2015 and 2016 shared tasks on Shallow Discourse Parsing (Xue et al., 2015, 2016) have set the standard to second-level classification. The second-level classification is more useful for most downstream tasks. Following other works we directly compare to in our evaluation, we here use the setting where AltLex, EntRel and NoRel tags are ignored. About 2.2% of the implicit relation instances in PDTB have been annotated with two relations, these are considered as two training instances.

To allow for full comparability to earlier work, we here report results for three different settings. The first one is denoted as PDTB-Lin (Lin et al., 2009); it uses sections 2-21 for training, 22 as dev and section 23 as test set. The second one is labeled PDTB-Ji (Ji and Eisenstein, 2015), and uses sections 2-20 for training, 0-1 as dev and evaluates on sections 21-22. Our third setting follows the recommendations of Chapter 3, and performs 10-fold cross validation on the whole corpus (sections 0-23). Table 6.1 shows the number of instances in train, development and test set in different settings.

As advocated in Chapter 3, cross validation approach addresses problems related to the small corpus size, and reports model performance across all folds. This to some extend avoids the risks of unreliable conclusions that are made on the small test set in the conventional data split settings.

Preprocessing

We first convert tokens in PDTB to lowercase and normalize strings, which removes special characters. The word embeddings used for initializing the word representations are trained with the CBOW architecture in *Word2Vec*[2] (Mikolov et al., 2013) on PDTB training set. All the weights in the model are initialized with uniform random.

[1] Cross-validation allows us to test on all 15057 instances.

[2] `https://code.google.com/archive/p/word2vec/`

To better locate the connective positions in the target side, we use two position indicators ($\langle conn \rangle$, $\langle /conn \rangle$) which specify the starting and ending of the connectives (Zhou et al., 2016), which also indicate the spans of discourse arguments.

Since our main task here is not generating arguments, it is better to have representations generated by correct words rather than by wrongly predicted ones. So at test time, instead of using the predicted word from previous time step as current input, we use the source sentence as the decoder's input and target. As the implicit connective is not available at test time, we use a random vector, which we used as "impl_conn" in Figure 6.2, as a placeholder to inform the sequence that the upcoming word should be a connective.

Hyper-parameters

There are several hyper-parameters in our model, including dimension of word vectors d, two dropout rates after embedding layer q_1 and before softmax layer q_2, two learning rates for encoder-decoder lr_1 and for classifier lr_2, top k for k-max pooling layer, different weights w for losses in Equation (6.17) and λ denotes the coefficient of regularizer, which controls the importance of the regularization term, as shown in Table 6.2.

d	q_1	q_2	lr_1	lr_2	k	w	λ
100	0.5	0.2	$2.5e^{-3}$	$5e^{-3}$	5	0.2	$5e^{-4}$

Table 6.2: Hyper-parameter settings.

6.4.2 Model training

To train our model, the training objective is defined by the loss function we introduced above. We use Adaptive Moment Estimation (Adam) (Kingma and Ba, 2014) with different learning rate for different parts of the model as our optimizer. Dropout layers are applied after the embedding layer and also on the top feature vector before the softmax layer in the classifier. We also employ L_2 regularization with small λ in our objective functions for preventing over-fitting. The values of the hyper-parameters, are provided in Table 6.2. The model is trained firstly to minimize the loss in Equation 6.15 until convergence, we use scheduled sampling (Bengio et al., 2015) during training to avoid "teacher-forcing problem".

Methods	PDTB-Lin	PDTB-Ji	Cross Validation
Majority class	26.11	26.18	25.59
Lin et al. (2009)	40.20	-	-
Qin et al. (2016a)	43.81	45.04	-
Cai and Zhao (2017)	-	45.81	-
Qin et al. (2017)	44.65	**46.23**	-
Shi et al. (2017) (with extra data)	**45.50**	-	**37.84**
Encoder only (Bi-LSTM) (Shi et al., 2017)	34.32	-	30.01
Auto-Encoder	43.86	45.43	39.50
Seq2Seq w/o Mem Net	45.75	47.05	40.29
Proposed Method	**45.82**	**47.83**	**41.29**

Table 6.3: Accuracy (%) of implicit discourse relations on PDTB-Lin, PDTB-Ji and Cross Validation Settings for multi-class classification.

And then to minimize the joint loss in Equation 6.17 to train the implicit discourse relation classifier.

6.4.3 Experimental results

Second-level multi-class classification

We compare our models with six previous methods, as shown in Table 6.3. The baselines contain feature-based methods (Lin et al., 2009), state-of-the-art neural networks at the time of the study (Qin et al., 2016a; Cai and Zhao, 2017), including the adversarial neural network that also exploits the annotated implicit connectives (Qin et al., 2017), as well as the data extension method based on using explicitated connectives from translation to other languages (as in Chapter 5).

Additionally, we ablate our model by taking out the prediction of the implicit connective in the sequence to sequence model. The resulting model is labeled Auto-Encoder in Table 6.3. And seq2seq network without knowledge memory, which means we use the output of gated interaction layer to predict the label directly, as denoted as Seq2Seq w/o Mem Net.

Our proposed model outperforms the other models in each of the settings. Compared with performances in Qin et al. (2017), although we share the similar idea of extracting highly discriminative features by generating connective-augmented representations for implicit dis-

course relations, our method improves about 1.2% on setting PDTB-Lin and 1.6% on the PDTB-Ji setting. The importance of the implicit connective is also illustrated by the fact that the "Auto-Encoder" model, which is identical to our model except it does not predict the implicit connective, performs worse than the model which does. This confirms our initial hypothesis that training with implicit connectives helps to expose the latent discriminative features in the relational arguments, and generates more refined semantic representation. It also means that, to some extent, purely increasing the size of tunable parameters is not always helpful in this task and trying to predict implicit connectives in the decoder does indeed help the model extract more discriminative features for this task. What's more, we can also see that without the memory network, the performances are also worse, it shows that with the concatenation of knowledge vector, the training instance may be capable of finding related instances to get common knowledge for predicting implicit relations. As Shi and Demberg (2017) argued that it is risky to conclude with testing on such small test set, we also run cross-validation on the whole PDTB. From Table 6.3, we have the same conclusion with the effectiveness of our method, which outperformed the baseline (Bi-LSTM) with more than 11% points and 3% compared with Shi et al. (2017) even though they have used a very large extra corpus.

For the sake of obtaining a better intuition on how the global attention works in our model, Figure 6.2 demonstrates the weights of different time-steps in attention layer from the decoder. The weights show how much importance the word attached to the source words while predicting target words. We can see that without the connective in the target side of test, the word filler still works as a connective to help predict the upcoming words. For instance, the true discourse relation for the right-hand example is *Expansion.Alternative*, at the word filler's time-step, it attached more importance on the negation "don't" and "tastefully appointed". It means the current representation could grasp the key information and try to focus on the important words to help with the task. Here we see plenty room for adapting this model to discourse connective prediction task, we would like to leave this to the future work.

We also try to figure out which instances' representations have been chosen from the memory matrix while predicting. Table 6.6 shows two examples and their context instances with

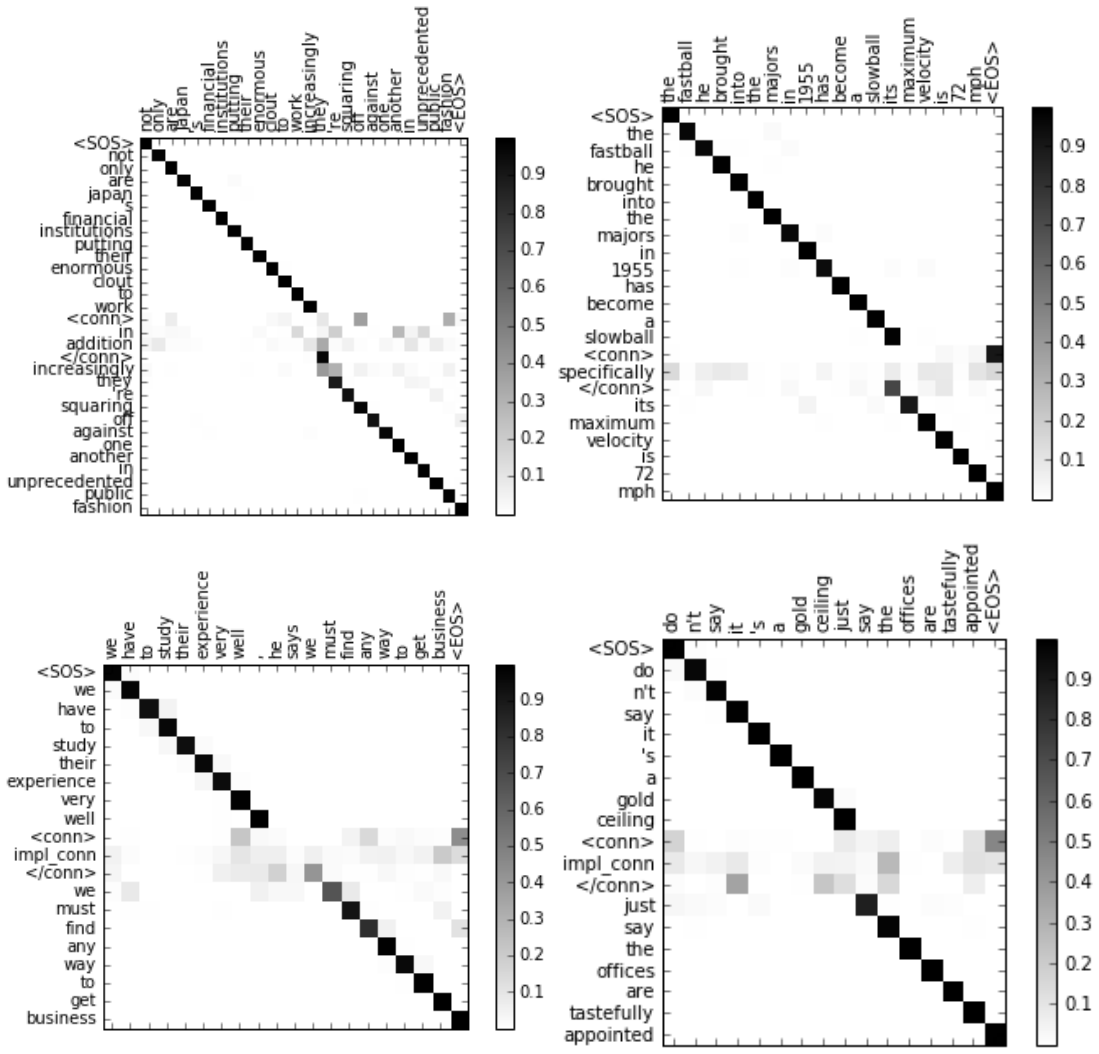

Figure 6.2: Visualization of attention weights during predicting target sentence in train and test, x-axis denotes the source sentence and the y-axis is the targets. Upper two figures are examples from training set with implicit connectives inside, while the bottom two, in which the implicit connectives have been replaced by the word filler "impl_conn", are from test set.

top 2 memory attentions among the whole training set. We can see that both examples show that the memory attention attached more importance on the same relations. This means that with the Context Memory, the model could facilitate the discourse relation prediction by choosing examples that share similar semantic representation and discourse relation during prediction.

Relation	Train	Dev	Test
Comparison	1855	189	145
Contingency	3235	281	273
Expansion	6673	638	538
Temporal	582	48	55
Total	12345	1156	1011

Table 6.4: Distribution of top-level implicit discourse relations in the PDTB.

Methods	Four-ways		One-Versus-all Binary (F_1)			
	F_1	Acc.	Comp.	Cont.	Expa.	Temp.
Rutherford and Xue (2014)	38.40	55.50	39.70	54.42	70.23	28.69
Qin et al. (2016b)	-	-	**41.55**	57.32	71.50	35.43
Liu et al. (2016)	44.98	57.27	37.91	55.88	69.97	37.17
Ji et al. (2016)	42.30	**59.50**	-	-	-	-
Liu and Li (2016)	46.29	57.17	36.70	54.48	70.43	**38.84**
Qin et al. (2017)	-	-	40.87	54.46	72.38	36.20
Lan et al. (2017)	**47.80**	57.39	40.73	**58.96**	**72.47**	38.50
Our method	46.40	**61.42**	**41.83**	**62.07**	69.58	35.72

Table 6.5: Comparison of F_1 scores (%) and Accuracy (%) with the State-of-the-art Approaches at the time of the study for four-ways and one-versus-all binary classification on PDTB. Comp., Cont., Expa. and Temp. stand for Comparison, Contingency, Expansion and Temporal respectively.

*In recent years, U.S. steelmakers have supplied about 80% of the 100 million tons of steel used annually by the nation. (**in addition,**) Of the remaining 20% needed, the steel-quota negotiations allocate about 15% to foreign suppliers.* — Expansion.Conjunction 1. The average debt of medical school graduates who borrowed to pay for their education jumped 10% to $42,374 this year from $38,489 in 1988, says the Association of American Medical Colleges. (**furthermore**) that's 115% more than in 1981 — Expansion.Conjunction 2. ... he rigged up an alarm system, including a portable beeper, to alert him when Sventek came on the line. (**and**) Some nights he slept under his desk. — Expansion.Conjunction
*Prices for capital equipment rose a hefty 1.1% in September, while prices for home electronic equipment fell 1.1%. (**Meanwhile,**) food prices declined 0.6%, after climbing 0.3% in August.* — Comparison.Contrast 1. Lloyd's overblown bureaucracy also hampers efforts to update marketing strategies. (**Although**) some underwriters have been pressing for years to tap the low-margin business by selling some policies directly to consumers. — Comparison.Contrast 2. Valley National "isn't out of the woods yet. (**Specifically**), the key will be whether Arizona real estate turns around or at least stabilizes — Expansion.Restatement

Table 6.6: Example of attention in Context Knowledge Memory. The sentences in italic are from PDTB test set and following 2 instances are the ones with top 2 attention weights from training set.

Top-level binary and 4-way classification

A lot of the recent works in PDTB relation recognition have focused on first level relations, both on binary and 4-ways classification. We also report the performance on level-one relation classification for more comparison to prior works. As described above, we followed the conventional experimental settings (Rutherford and Xue, 2015; Liu and Li, 2016) as closely as possible. Table 6.4 shows the distribution of top-level implicit discourse relation in PDTB, it's worth noticing that there are only 55 instances for Temporal Relation in the test set.

To make the results comparable with previous work, we report the F_1 score for four binary classifications and both F_1 and Accuracy for 4-way classification, which can be found in Table 6.5. We can see that our method outperforms all alternatives on COMPARISON and CONTINGENCY, and obtain comparable scores with the state-of-the-art in others at the time of the study. For 4-way classification, we got the best accuracy and second-best F_1 with around 2% better than in Ji et al. (2016).

6.5 Analysis and discussion

We present in this chapter a novel neural method trying to integrate implicit connectives into the representation of implicit discourse relations with a joint learning framework of sequence-to-sequence network. We conduct experiments with different settings on PDTB benchmark, the results show that our proposed method achieves state-of-the-art performance (at the time of publication) on recognizing the implicit discourse relations at the time of the study and the improvements are not only brought by the increasing number of parameters. The model also has great potential abilities in implicit connective prediction in the future.

Our proposed method shares similar spirit with previous work in Zhou et al. (2010), who also tried to leverage implicit connectives to help extract discriminative features from implicit discourse instances. Comparing with the adversarial method proposed by Qin et al. (2017), our proposed model more closely mimics humans' annotation process of implicit discourse relations and is trained to directly explicitate the implicit relations before classification. With the representation of the original implicit sentence and the explicitated one

from decoder, and the help of the explicit knowledge vector from memory network, the implicit relation could be classified with higher accuracy.

Although our method has not been trained as a generative model in our experiments, we can see potential for applying it to generative tasks. With more annotated data, minor modification and fine-tuned training, we believe our proposed method could also be applied to tasks like implicit discourse connective prediction, or argument generation in the future.

6.6 Summary

In this chapter, I propose a sequence-to-sequence based neural network which not only predicts the implicit discourse relation between the arguments but also has a secondary task to be trained to explicitate the discourse relation while generating the arguments at the same time. The experimental results show that with the secondary task and the memory network component, the proposed method achieved the best performances at the time of the study, on different settings including the first level four-ways, one-versus-all and the 11 ways classification on the second level relations.

Nonetheless, having good encoding only does part of the job. A good implicit discourse relation classification should be able to encode discourse expectation and learn typical temporal sequences, causes, consequences etc. for all kinds of events. In the next chapter, I try to figure out whether having the correct next sentence is beneficial for the task, with the help of the recently proposed model BERT (Devlin et al., 2019), which has a next sentence prediction as a subtask in the pre-training of the language model. And also try to understand how much the pre-trained language model helps when it meets with a new domain.

Chapter 7

Next Sentence Prediction helps within and across domains

In the last chapters, we prove that having better understanding, encoding and semantic interpretation are vital and beneficial to the task of implicit discourse relation classification. However, taking the context to a broader range, it only does part of the work. A good implicit discourse relation classifier should also be aware of the upcoming events, causes, consequences etc. to encode the discourse expectations into the sentence representations. In this chapter, we try to figure out whether having the correct next sentence is beneficial for the task or not, with the help of the recently proposed BERT model.

7.1 Introduction

Discourse relations in texts are sometimes marked with an explicit connective (e.g., *but*, *because*, *however*), but these explicit signals are often absent. When there is no connective, classification has to rely on semantic information from the relational arguments. This task is very challenging, with the state-of-the-art systems at the time of the study achieving accuracy of only 45% to 48% on 11-way classification. Let's consider example 7.1:

(1) [*The joint venture with Mr. Lang wasn't a good one.*]$_{Arg1}$ [**The venture, formed in 1986, was supposed to be Time's low-cost, safe entry into women's magazines.**]$_{Arg2}$
implicit `Comparison.Concession.Expectation` relation from PDTB: wsj_1903

In order to correctly classify the relation, it is necessary to understand that Arg1 raises the expectation that the next discourse segment may provide an explanation for why the venture wasn't good (e.g., that it was risky), and Arg2 contrasts with this discourse expectation. More generally, this means that a successful discourse relation classification model would have to be able to learn typical temporal event sequences, reasons, consequences etc. for all kinds of events. Statistical models attempted to address this intuition by giving models word pairs from the two arguments as features (Lin et al., 2009; Park and Cardie, 2012; Biran and McKeown, 2013; Rutherford and Xue, 2014), so that models could for instance learn to recognize antonym relations between words in the two arguments.

Recent models exploit such similarity relations between the two arguments, as well as simpler surface features that occur in one relational argument and correlate with specific coherence relations (e.g., the presence of negation, temporal expressions etc. may give hints as to what coherence relation may be present, see Park and Cardie (2012); Asr and Demberg (2015)). However, relations between arguments are often a lot more diverse than simple contrasts that can be captured through antonyms, and may rely on world knowledge (Kishimoto et al., 2018). It is hence clear that one cannot learn all these diverse relations from the very small amounts of available training data. Instead, we would have to learn a more general representation of discourse expectations.

In fact, most of today's discourse relation classifiers attempt to predict the coherence relation between relational arguments, without having access to the world knowledge that is often necessary to assess how two events relate to one another. Instead, they capture some generalizable patterns between word pairs in the relational arguments, e.g. if the relational arguments contain antonyms, it's highly possible to be a contrastive relation. And the largest annotated discourse relation corpus, describes a text as a series of discourse relations, each of which consists of two arguments and a connective. The discourse relation can either be *explicit* or *implicit* depending on whether the connective is presence or not.

Many recent discourse relation classification approaches have focused on cross-lingual data

augmentation (Chapter 4, 5), training models to better represent the relational arguments by using various neural network models, including feed-forward network (Rutherford et al., 2017b), convolutional neural networks (Zhang et al., 2015), recurrent neural network (Ji et al., 2016; Bai and Zhao, 2018), character-based (Qin et al., 2016a) or formulating relation classification as an adversarial task (Qin et al., 2017). These models typically use pre-trained semantic embeddings generated from language modeling tasks, like Word2Vec (Mikolov et al., 2013), GloVe (Pennington et al., 2014) and ELMo (Peters et al., 2018).

However, previously proposed neural models still crucially lack a representation of the *typical relations between sentences*: to solve the task properly, a model should ideally be able to form discourse expectations, i.e., to represent the typical causes, consequences, next events or contrasts to a given event described in one relational argument, and then assess the content of the second relational argument with respect to these expectations (see Example 7.1). Previous models would have to learn these relations only from the annotated training data, which is much too sparse for learning all possible relations between all events, states or claims.

The recently proposed BERT model (Devlin et al., 2019) takes a promising step towards addressing this problem: the BERT representations are trained using a language modelling and, crucially, a "next sentence prediction" task, where the model is presented with the actual next sentence vs. a different sentence and needs to select the original next sentence. We believe it is a good fit for discourse relation recognition, since the task allows the model to represent what a typical next sentence would look like.

In this chapter, we show that a BERT-based model outperforms the current state of the art by 8% points in 11-way implicit discourse relation classification on PDTB. We also show that after pre-training with small size cross-domain data, the model can be easily transferred to a new domain: it achieves around 16% accuracy gain on BioDRB compared to state of the art model. We also show that the Next Sentence Prediction task played an important role in these improvements.

7.2 BERT

Devlin et al. (2019) proposed the Bidirectional Encoder Representation from Transformers (BERT), which is designed to pre-train a deep bidirectional representation by jointly conditioning on both left and right contexts. BERT is trained using two novel unsupervised prediction tasks: Masked Language Modeling and Next Sentence Prediction (NSP).

7.2.1 Masked language model

Given that both the contexts on the left and right side are vital for the language models, it is reasonable to believe that a deep bidirectional model is more powerful than either a left-to-right or a shallow concatenation of a left-to right and right-to-left model. In other words, a deep bidirectional language model should be able to encode the surrounding contexts at the same time. In this way, each word is able to indirectly see itself and the model could trivially predict the target word in a multi-layered context.

Devlin et al. (2019) proposed a masked language model, in which some of the input tokens are randomly chosen and masked and the final hidden vectors corresponding to the mask tokens are fed into an output softmax over the vocabulary, as in a standard language model. Specifically, 15% of all tokens at random in each sequence are chosen to be replaced by (1) the actual "[MASK]" token 80% of the time. (2) a random token 10% of the time. (3) the unchanged token 10% of the time. This is different compared with the denoising auto-encoders such as sequence-to-sequence models, the masked language model only predicts the masked words rather than reconstructing the entire input recurrently.

7.2.2 Next sentence prediction

Many downstream NLP tasks such as Question Answering (QA) and Natural Language Inference (NLI) are based on understanding the relationship between two sentences, which is not directly captured by the language model. The next sentence prediction task has been formulated as a binary classification task: the model is trained to distinguish the originally following sentence from a randomly chosen sentence from the corpus, and it showed great

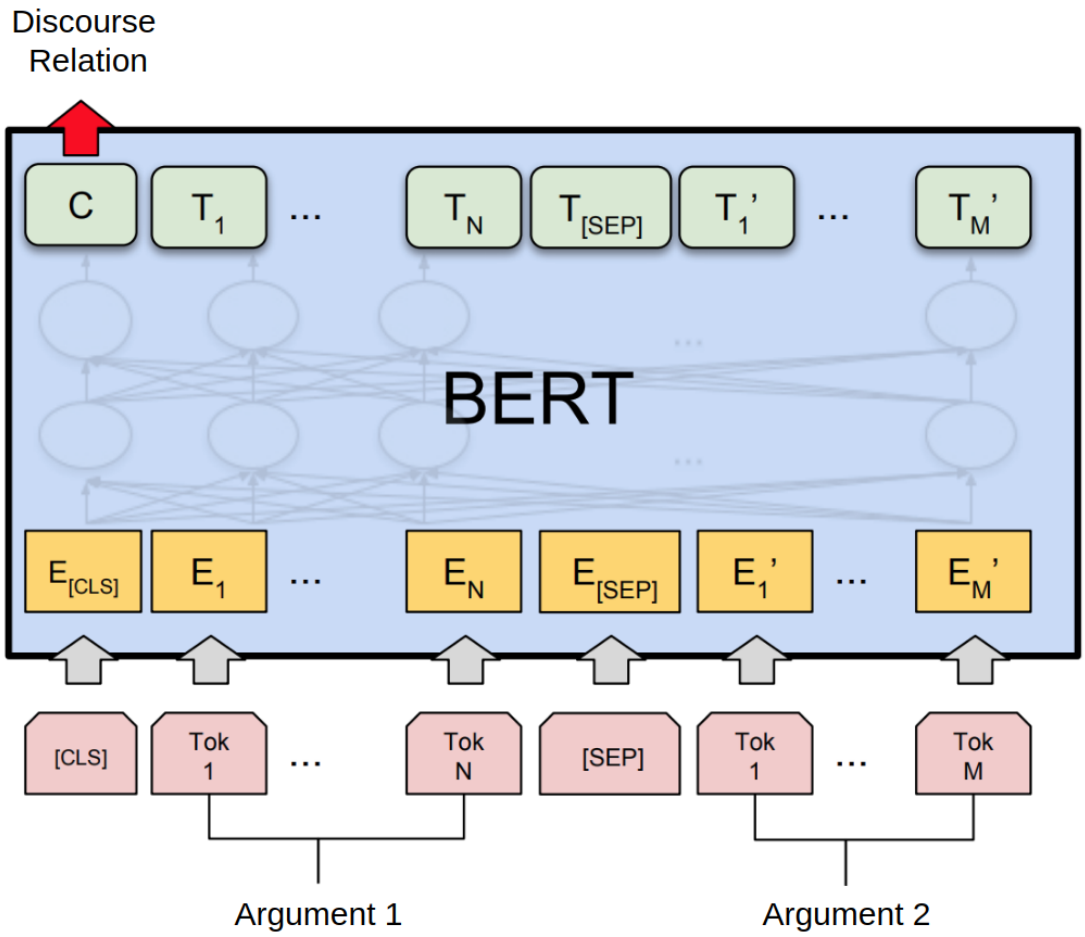

Figure 7.1: The architecture from BERT (Devlin et al., 2019) for fine-tuning of implicit discourse relation classification.

helps in multiple NLP tasks especially inference ones. In more detail, the model is pre-trained for a binarized *next sentence prediction* task that can be trivially generated from any monolingual corpus. Specifically, when choosing the sentence A and B for each training example, 50% of the time B is the actual next sentence that follows A, and 50% of the time it is a random sentence from the corpus. The resulting BERT representations thus encode a representation of upcoming discourse content, and hence contain discourse expectation representations which, as we argued above, are required for classifying coherence relations.

Methods	PDTB-Lin	PDTB-Ji	Cross Validation
Cai and Zhao (2017)	-	45.81	-
Kishimoto et al. (2018)	38.77	-	39.80
Bai and Zhao (2018)	45.73	48.22	-
Shi and Demberg (2019a)	45.82	47.83	41.29
Bi-LSTM + w2v_300	37.95(0.59)	40.57(0.67)	37.82(0.74)
BERT	**53.13(0.37)**	**53.30(0.39)**	**49.30(1.33)**
BERT + WSJ w/o NSP	**53.39(0.49)**	**51.28(0.49)**	**49.32(1.24)**
BERT + WSJ	**54.82(0.61)**$*$	**53.23(0.32)**$*$	**49.35(0.83)**

Table 7.1: Accuracy (%) with standard deviation in brackets of implicit discourse relation classification on different settings of PDTB level 2 relations. NSP refers to the subtask "next sentence prediction" in the pre-training of BERT. Numbers in bold signal significant improvements over the previous state of the art ($p<0.01$). Numbers with $*$ denote significant improvements over BERT + WSJ w/o NSP with $p<0.01$.

7.3 Experiments and results

As shown in Figure 7.1, *E* denotes the input tokens' embedding and *T* are the target words. In our case they are identical. We fit the implicit discourse relation task into sentence-pair classification proposed in BERT. Argument 1 and Argument 2 are separated with token "[SEP]"; "[CLS]" is the special classification embedding while "C" is the same as "[CLS]" in pre-training but the ground-truth label in the fine-tuning. In the experiments, we used the uncased base model[1] provided by Devlin et al. (2019), which is trained on *BooksCorpus* and *English Wikipedia* with 3300M tokens in total.

7.3.1 On PDTB

We use the Penn Discourse Tree Bank (Prasad et al., 2008), the largest available manually annotated discourse corpus. It provides a three level hierarchy of relation tags. Following the experimental settings and evaluation metrics in Bai and Zhao (2018), we use two most-used splitting methods of PDTB data, denoted as PDTB-Lin (Lin et al., 2009), which uses sections 2-21, 22, 23 as training, validation and test sets, and PDTB-Ji (Ji and Eisenstein, 2015), which uses 2-20, 0-1, 21-22 as training, validation and test sets and report the overall

[1] `https://github.com/google-research/bert\#pre-trained-models`

accuracy score. In addition, we also performe 10-fold cross validation among sections 0-22, as promoted in Shi and Demberg (2017). We also follow the standard in the literature to formulate the task as an 11-way classification task.

Results are presented in Table 7.1. We evaluate three versions of the BERT-based model. All of our BERT models use the pre-trained representations and are fine-tuned on the PDTB training data. The version marked as "BERT" does not do any additional pre-training. BERT+WSJ in addition performs further pre-training on the parts of the *Wall Street Journal* corpus that do not have discourse relation annotation. The model version "BERT+WJS w/o NSP" also performs pre-training on the WSJ corpus, but only uses the Masked Language Modelling task, not the Next Sentence Prediction task in the pre-training. We added this variant to measure the benefit of in-domain NSP on discourse relation classification (note though that the downloaded pre-trained BERT model contains the NSP task in the original pre-training).

We compare the results with four state-of-the-art systems: Cai and Zhao (2017) proposed a model that takes a step towards calculating discourse expectations by using attention over an encoding of the first argument, to generate the representation of the second argument, and then learning a classifier based on the concatenation of the encodings of the two discourse relation arguments. Kishimoto et al. (2018) fed external world knowledge (ConceptNet relations and coreferences) explicitly into MAGE-GRU (Dhingra et al., 2017) and achieved improvements compared to only using the relational arguments. However, we here show that it works even better when we learn this knowledge implicit through next sentence prediction task. Shi and Demberg (2019a) used a seq2seq model that learns better argument representations due to being trained to explicitate the implicit connective. In addition, their classifier also uses a memory network that is intended to help remember similar argument pairs encountered during training. The current best performance was achieved by Bai and Zhao (2018), who combined representations from different grained embeddings including contextualized word vectors from ELMo (Peters et al., 2018), which has been proved very helpful. In addition, we compare our results with a simple bidirectional LSTM network and pre-trained word embeddings from Word2Vec.

Method	Cross-Domain	In-Domain
Bi-LSTM + w2v_300	32.97	46.49
Bai and Zhao (2018)	29.52	55.90
BioBERT (Lee et al., 2020)	**44.33**	**67.58**
BERT	**44.79**	**63.02**
BERT + GENIA w/o NSP	**43.99**	**65.02**
BERT + GENIA	**45.19**$*$	**66.04**$*$

Table 7.2: Accuracy (%) on BioDRB level 2 relations with different settings. Cross-Domain means trained on PDTB and tested on BioDRB. For the In-Domain setting, we used 5-fold cross-validation and report average accuracy. Numbers in bold are significantly better than the state of the art system with $p<0.01$ and numbers with $*$ denote denote significant improvements over BERT + GENIA w/o NSP with $p<0.01$.

7.3.2 On BioDRB

The Biomedical Discourse Relation Bank (Prasad et al., 2011) also follows PDTB-style annotation. It is a corpus annotated over 24 open access full-text articles from the GENIA corpus (Kim et al., 2003) in the biomedical domain. Compared with PDTB, some new discourse relations and changes have been introduced in the annotation of BioDRB. In order to make the results comparable, we preprocess the BioDRB annotations to map the relations to the PDTB ones, following the instructions in Prasad et al. (2011).

The biomedical domain is very different from the WSJ or the data on which the BERT model was trained. The BioDRB contains a lot of professional words / phrases that are extremely hard to model. In order to test the ability of the BERT model on cross-domain data, we performed fine-tuning on PDTB while testing on BioDRB. We also tested the state of the art model of implicit discourse relation classification proposed by Bai and Zhao (2018) on BioDRB. From Table 7.2, we can see that the BERT base model achieved almost 12% points improvement over the Bi-LSTM baseline and 15% points over Bai and Zhao (2018). When fine-tuned on in-domain data in the cross-validation setting, the improvement increases to around 17% points.

It is also interesting to know whether the performance of the BERT model can be improved if we add additional pre-training on in-domain data. BioBert (Lee et al., 2020) continues pre-training BERT with bio-medical texts including PubMed and PMC corpora (around 18B

Method	Comp.	Cont.	Exp.	Temp.	Average
Naïve Bayes (Xu et al., 2012)	7.69(61.45)	3.88(60.24)	68.54(55.02)	17.19(57.43)	24.33(58.54)
MaxEnt (Xu et al., 2012)	7.08(57.83)	3.36(53.82)	72.32(60.64)	23.81(61.45)	26.64(58.44)
BERT	**8.25(63.82)**	**10.26(71.54)**	**90.20(86.18)**	**63.41(87.80)**	**43.03(77.34)**

Table 7.3: F_1-score (Accuracy) of binary classification on level 1 implicit relation in BioDRB.

tokens), which achieved the best results on in-domain setting. Similarly, BERT+GENIA refers to a model in which the downloaded BERT representations are further pre-trained on the parts of the GENIA corpus which consists of 18k sentences and is not annotated with coherence relations. Evaluation shows that this in-domain pre-training yields another 3% point improvement; our tests also show that the NSP task again plays a substantial role in the improvement. We believe that gains for further pre-training on GENIA for the biomedical domain are higher than for pre-training on WSJ for PDTB because the domain difference between the *BooksCorpus* and the biomedical domain is larger.

Currently there are not so many published results that we can compare with on BioDRB for implicit discourse relation classification. We compare BERT model with naïve Bayes and MaxEnt methods proposed in Xu et al. (2012) on one-versus-all binary classification. We followe the settings in Xu et al. (2012) and used two articles ("GENIA_1421503", "GENIA_1513057") for testing and one article ("GENIA_111020") for validation. During training, we employ down-sampling or up-sampling to keep the numbers of positive and negative samples in each relation consistent. The BERT base model achieves 43.03% average F_1 score and 77.34% average accuracy in one-versus-all level-1 classification. Compared with the current state-of-the-art performances (26.64% F_1 and 58.54% accuracy) in Xu et al. (2012), it achieves around 16% and 19% points improvement when trained in-domain, as illustrated in Table 7.3.

7.4 Conclusion and discussion

The usage of the BERT model in this chapter is motivated primarily by the use of the next-sentence prediction task during training. The results in Table 7.1 and Table 7.2 confirm that removing the "Next Sentence Prediction" hurts the performance on both PDTB and

Relations	WSJ w/o NSP			WSJ w/ NSP			Count
	P	R	F_1	P	R	F_1	
Temporal.Asynchronous	0.38	0.46	0.41	0.29	0.38	0.33	13
Temporal.Synchrony	-	-	-	-	-	-	5
Contingency.Cause	0.57	0.65	0.61	0.57	0.64	0.60	200
Contingency.Pragmatic Cause	-	-	-	-	-	-	5
Comparison.Contrast	0.55	0.48	0.51	0.54	0.57	0.55	127
Comparison.Concession	-	-	-	-	-	-	5
Expansion.Conjunction	0.42	0.60	0.49	0.46	0.61	0.53	118
Expansion.Instantiation	0.62	0.67	0.64	0.62	0.65	0.64	72
Expansion.Restatement	0.52	0.45	0.48	0.55	0.45	0.50	190
Expansion.Alternative	0.83	0.33	0.48	0.67	0.40	0.50	15
Expansion.List	0.71	0.17	0.27	0.60	0.20	0.30	30
Micro Avg.	0.53	0.53	0.52	0.55	0.55	0.55	780

Relations	GENIA w/o NSP			GENIA w/ NSP			Count
	P	R	F_1	P	R	F_1	
Temporal.Asynchronous	-	-	-	-	-	-	-
Temporal.Synchrony	0.87	0.84	0.85	0.90	0.88	0.89	80
Contingency.Cause	0.22	0.10	0.14	0.23	0.15	0.18	20
Contingency.Pragmatic Cause	-	-	-	-	-	-	1
Comparison.Contrast	-	-	-	-	-	-	22
Comparison.Concession	-	-	-	0.50	0.06	0.11	16
Expansion.Conjunction	0.60	0.78	0.68	0.62	0.82	0.71	130
Expansion.Instantiation	-	-	-	-	-	-	9
Expansion.Restatement	0.56	0.76	0.65	0.59	0.69	0.64	72
Expansion.Alternative	-	-	-	-	-	-	1
Expansion.List	-	-	-	-	-	-	-
Micro Avg.	0.55	0.64	0.59	0.59	0.66	0.61	351

Table 7.4: Precision, Recall and F_1 score for each level-2 relation on PDTB-Lin setting and BioDRB with "BERT + WSJ/GENIA" systems w/ and w/o NSP. "-" indicates 0.00 and "C." means the number of each relation in the test set.

BioDRB.

In order to have better insights about which relation has benefited from the NSP task, we also report the detailed performance for each relation with and without it in BERT. As illustrated in Table 7.4, we can see that performances on relations like *Temporal.Synchrony, Comparison.Contrast, Expansion.Conjunction and Expansion.Alternative* have been improved by a large margin. This shows that representing the likely upcoming sentence helps the model form discourse expectations, which the classifier can then use to predict the coherence relation between the actually observed arguments.

However, compared with BERT+GENIA, the results of BioBert (Lee et al., 2020) in Table 7.2 show that having large in-domain data for pre-training also has limited ability in learning domain specific representations. We therefore believe that the model could be further improved by including external domain-specific knowledge from an ontology (as in Kishimoto et al. (2018)) or a causal graph for biomedical concepts and events.

7.5 Summary

In this chapter, we show that BERT has very good ability in encoding the semantic relationship between sentences with its "next sentence prediction" task in pre-training. It outperforms the current state-of-the-art systems significantly with a substantial margin on both in-domain and cross domain data. Our results also indicate that the next-sentence prediction task during training indeed plays a role in this improvement.

However, the performance of BioBERT shows limited ability in learning domain specific representations. In the next chapter, I will explore the joint representation of discourse expectations through implicit representations that are learned during training and the inclusion of external domain-specific knowledge. In addition, Yang et al. (2019) shows that NSP only helps tasks with longer texts. It would be interesting to see whether it has the same effect on implicit discourse relation classification task, we'd like to leave that in the future work.

Chapter 8

Entity Enhancement for Implicit Discourse Relation Classification

8.1 Introduction

Natural Language Processing (NLP) systems often perform below what is expected or observed during the experimental testing, when they are being used in a real-life application. One of the most important reasons is that when training the system, a huge assumption has been made that the training and testing data come from a common underlying distribution (Li, 2012). However, oftentimes the training data is too specialized to provide a generalized estimation about the distribution. This problem leads to one of the core challenges in designing a common computational language understanding system, how to adapt the trained system to a new domain. For example, in the part-of-speech tagging task, the word *monitor* is likely to be a verb in the financial domain, like the **Wall Street Journal (WSJ)** corpora, while in the corpus of computer hardware, it's more likely to be a noun. Therefore domain adaptation algorithms are designed to bridge the distribution gap between the training data and the test data.

As we discussed in the previous chapter, the Penn Discourse Treebank (PDTB) and Biomed-

ical Discourse Relation Bank (BioDRB) are the most popular corpora for implicit discourse relation classification. The domains' gap (financial vs. biomedical) makes it very difficult to learn general information from each other. Even with the powerful BERT model (Devlin et al., 2019), the performance trained cross domain is around 20% accuracy worse than the in-domain one.

The target of domain adaptation is to bring the underlying probability in the source domain $p_s(x, y)$ closer to the target domain $p_t(x, y)$. It basically can be categorized into three categories (Li, 2012):

Instance-based methods: This set of methods revolve around selecting and/or weighting instances in the source domain that are similar to those in the target domain. This type of algorithm is closely related to common semi-supervised learning framework such as self-training. The general idea is by weighting and selecting training instances, the influences from the empirical distribution can be recovered. The main challenge in this class is how to determine the instance weighting parameters or select which instance for training.

Feature transformation based methods: in this class, the assumption is that $p_s(x, y)$ differs from $p_t(x, y)$, but there exist some general features $x_g \in \chi$ that identical or similar conditional distribution in both the source and target domains. In this case, there are some correspondence between $p_s(x, y)$ and $p_t(x, y)$, which means it's possible to project the original feature space χ into a new space χ_t by using some projection methods. For instance, as shown in the Table 8.1, for the part-of-speech tagging task, some words differ in two domains either because they have different meanings in the two domains or one feature occurs in one domain but rarely occurs in another. For example, *investment* only occurs in the *Wall Street Journal*, while the verb *required*, the prepositions *from* and *for* have same meanings in the two domains. There are two main challenges in this type of methods: (1) How to distinguish domain-specific features and general features. (2) How to find the projection

Biomedical	Wall Street Journal
the **signal** *required* to	**investment** *required*
stimulatory **signal** *from*	**buyouts** *from* buyers
essential **signal** *for*	to **jail** *for* violating

Table 8.1: Example of general and domain-specific features (Blitzer et al., 2006). The italicized words are general features and bold are domain-specific.

from source domain to the target. To address these two questions, previous work proposed *structural correspondence learning* (Blitzer et al., 2006) and *Topic modeling* (Guo et al., 2009; Xue et al., 2008).

Prior-based methods: these methods place different priors over the parameters or the labels to make the estimated $p_s(y|x;\theta)$ closer to $p_t(y|x)$. This kind of method exploits model priors to overcome the distribution gap between two domains. One assumption is that the conditional distribution $p(x|y)$ is the same or similar in both domains, and that the gap between posterior distributions mainly come from the different priors between $p_s(y)$ and $p_t(y)$. Therefore a good estimation $p_t(y)$ based on the available data can boost the performance with a generative model such as the Naive Bayes framework.

Apart from the above methods, transfer learning techniques with neural networks have seen great successes in recent years. Pre-trained representations from deep neural network such as convolutional neural network, recurrent neural network and transformers have worked extremely well across a wide array of tasks no matter in computer vision, speech recognition or natural language processing.

Domain-specific embeddings: Mikolov et al. (2013) proposed Word2Vec to use only the words within an n-size window of a target word, and employed tricks such as negative sampling and hierarchical softmax layer to reduce computation time. GloVe (Pennington et al., 2014) used global co-occurrence information rather than just the local context used by Word2Vec, and train a regression model to minimize the difference between the vector dot product of two words and the log of their co-occurrence ratio. Until now, pre-trained word2vec embedding remains the most popular word representations and is mostly used as the initial embedding layer of a task-specific network. However, without special supervision on the domains, they still perform poor across domains. Bollegala et al. (2015) introduce the cross-domain word representation task, where the goal is to learn a domain-specific representation for each common word w. They constrain the representation of pivot features to be similar across domains by predicting non-pivot features from the surrounding pivot features.

Let's have a deeper look into an example from the Biomedical Discourse Relation Bank (BioDRB), as shown below. It would be very difficult for language models to learn domain-

specific embeddings for professional and rare entities like the bold entity words in the example. Given that the entities play an important role in inferring the implicit discourse relation, having an emphasis on entities when learning with pre-trained language model seem vital for implicit discourse relation classification.

1. *[The synovial membrane of **rheumatoid arthritis** (RA) is characterized by an infiltrate of a variety of inflammatory cells, such as **lymphocytes**, **macrophages**, and **dendritic cells**, together with proliferation of synovial fibroblast-like cells.]*$_{Arg1}$ *(**Implicit=As a result,**)* *[Numerous **cytokines** are overproduced in the inflamed joint.]*$_{Arg2}$

—Implicit, Contingency.Cause

In the last chapter, we show that the BioBERT (Lee et al., 2020), which is continuously pre-trained with BERT on bio-medical texts including PubMed and PMC corpora (around 18 billion tokens), still has a very limited ability for learning domain specific knowledge, i.e. entity information, entity relations etc.. This means that integrating external domain-specific knowledge may be beneficial for this task, which also has been found by Kishimoto et al. (2018), who integrated the ConceptNet relations as additional knowledge into the LSTM network and achieved better performance on the PDTB. Thus in this chapter, we first propose an unsupervised method using information retrieval and knowledge graph techniques with the assumption that if two instances share the same entities in both the relational arguments, there are high possibilities that they have the same or similar discourse relation. We then proceed to use the extracted relevant entities to enhance the pre-trained model to help better encode the meaning of the arguments. Comparing with the work in Kishimoto et al. (2018), we use the information retrieval system to get more topic and domain specific knowledge rather than the general knowledge base, and is not constrained by the pre-defined relations in the knowledge base such as ConceptNet etc.

8.2 Unsupervised methods with information retrieval system

Sparse data like BioDRB, which have only around 2,000 labeled implicit instances in total, are hard to encode. It is essential to use relevant similar explicit instances to help find the latent patterns they share. In this section, we introduce an unsupervised method for implicit discourse relation classification with the help of an information retrieval system and a knowledge graph. The motivation here is that if the similar entities appear at the different discourse instances, it is possible that the instances share some latent patterns in terms of the discourse relation. Since that explicit instances are relatively easy to identify by the discourse parser with high accuracy (namely 96%), with a large amount of unlabelled raw data, we believe that the system is able to identify those patterns and have statistically reasonable predictions with the help of the extracted corresponding explicit discourse instances.

We first identify the relevant documents and with a discourse parser, it is easy to get the explicit instances by identifying the discourse markers. Then we extract SPO (subject, predicate and Object) triples with the knowledge graph system. After matching the entities in the explicit instances and those in the query, we get different labels from the explicit instances that have similar entities in them. Finally we label the query with the majority vote.

8.2.1 Overview of the proposed method

Figure 8.1 illustrates the overall pipeline of the proposed method. First, each instance from BioDRB (Prasad et al., 2011) is seen as a query and fed into the PubMed [1] and PMC [2] database.

PubMed and **PMC** are free full-text archive of biomedical and life sciences journal literature at the U.S. National Institute of Health's National Library of Medicine. The database we use here is a subset of the whole PubMed and PMC collections. It consists of 7079 documents in total (1,376 for pubMed and 5,703 for PMC).

[1] PubMed [Internet]. Bethesda (MD): National Library of Medicine (US). [1946]. Available from: `https://www.ncbi.nlm.nih.gov/pubmed/`

[2] PubMed Central (PMC) [Internet]. Bethesda (MD): National Library of Medicine (US), National Center for Biotechnology Information; 2000. Available from: `https://www.ncbi.nlm.nih.gov/pmc/`

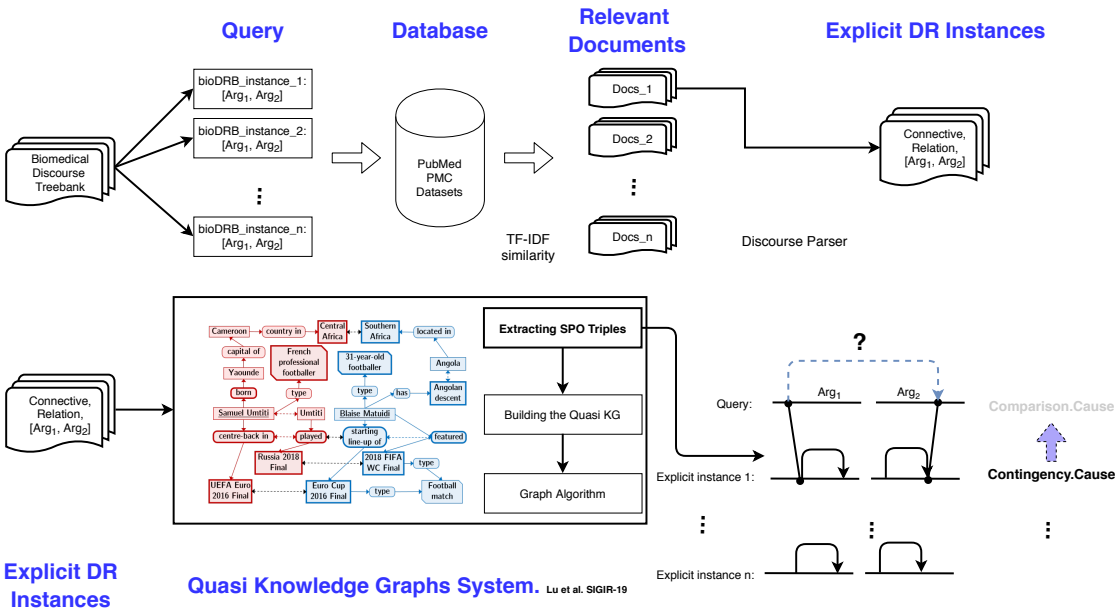

Figure 8.1: The Pipeline of the Proposed Method.

With the query and candidate documents, we employ the Term Frequency-inverse Document Frequency (TF-IDF) with the equation below to extract the top 10 relevant documents.

$$
\begin{aligned}
tfidf(t,d,D) &= tf(t,d) \cdot idf(t,D) \\
tf(t,d) &= \frac{f_{t,d}}{\sum_{t' \in d} f_{t',d}} \\
idf(t,D) &= \log \frac{N}{|t \in D : t \in d|}
\end{aligned}
\tag{8.1}
$$

with

- $f_{t,d}$: the raw count of term t, it equas to 1 if term t appears in document d, otherwise 0.
- N: total number of documents in the corpus N = |D|
- $|d \in D : t \in d|$: number of documents where the term t appears.

The candidate documents are later fed into the discourse parser, here we use the PDTB-style

end-to-end parser by Lin et al. (2014). The outputs of the parser contain the two arguments, discourse relation and the explicit discourse marker.

The **Quasi Knowledge Graphs System** proposed by Lu et al. (2019) is designed to answer complex questions. It is a novel method that computes direct answers to complex questions by dynamically tapping arbitrary text sources and joining sub-results from multiple documents, combining the versatility of text-based QA with graph-structure awareness of knowledge graph. It consists of several steps including graph constructing, building the quasi knowledge graph and graph algorithm. As we only need to get the SPO triples and have no questions to answer, next we will briefly introduce the first step in the graph constructing which is extracting SPO triples.

In order to identify cues in the matching documents and to join them across multiple documents, Lu et al. (2019) apply OpenIE (Kadry and Dietz, 2017; Mausam, 2016) to extract SPO triples. Popular tools for Open IE include OpenIE (Angeli et al., 2015), OpenIE 5.0 (Mausam, 2016), and ClausIE (Del Corro and Gemulla, 2013). However, each of these tools comes with limitations, such as that Stanford OpenIE focuses on precision, and produces correct but relatively few triples; OpenIE 5.0 and ClausIE often produce very long objects from complex sentences and make it difficult to align them across different triples and rendering subsequent graph construction infeasible. Therefore, Lu et al. (2019) devised their IE method with judicious consideration to phrase ordering and term proximities, which is also a very good fit for our task since we want that there is only one verb or noun phrase between the subject and object and don't have any other misleading graph edges. What's more, discourse relations are also very sensitive to the phrase ordering, for example, the reason and result can be either in the argument 1 or 2 respectively and leads to different discourse relations.

Firstly, they extracted the following triples respecting phrase ordering: Verb-phrase-mediated triples and Noun-phrase-mediated triples. And then they used pairwise distances between the triples' parts (S, P or O) in the document where it stems from. They define the distance d between two items as the number of intruding words plus one, and the score is set to $1/d$. This captures the intuition that the closer two parts are in text, the higher their score is. In this way, each of the SPO triples is associated with a confidence score.

After extracting the SPO triples from all the explicit discourse instances, we employ two types of matching strategy to connect them with the query: (i) **Hard matching**, which means that if the subject and object appear in the query respectively or vice versa, we count it as a vote. (ii) **Soft matching**. We find that with the hard matching, lots of positive samples have been filtered out and very few explicit instances stay. Therefore, we use the cosine distance between the subject / object and the noun phrases in the query, to measure the similarities. To be specific, we use the pre-trained BioBERT to encode the noun phrases. With a threshold, we can define whether they are close enough to make the corresponding explicit instance to be counted as a valid vote or not. This way is also used in the later entity-enhanced method with the pre-trained language model K-BERT (Liu et al., 2020).

With the steps described above, eventually each query has been connected to a number of similar explicit instances and the prediction for the query is the majority vote from all of them with their explicit discourse sense labels.

8.2.2 Experiments and results

Given that we haven't used any of the labels of the instances in the BioDRB as supervision for the training, we therefore can use all the instances as test sets. The experimental results are shown in Table 8.2. We compare the results with some of the related work that have results across domains.

- Bai and Zhao (2018) combines representations from different levels of embeddings including character-based embeddings, subword embeddings, and contextualized word vectors from ELMo.
- Bi-LSTM + Word2Vec: The simple bidirectional LSTM network with the 300 dimension word embeddings.
- BERT: The uncased base model with 12 layers and 768 hidden dimensions.
- BERT + GENIA: Continue training BERT with in-domain raw texts GENIA. It is the same as in the chapter 7.
- BioBERT (Lee et al., 2020): Pre-trained Bert with 18B in-domain tokens.

Methods	Cross-domain
Bai and Zhao (2018)	29.52
Bi-LSTM + Word2Vec	32.97
BERT (Devlin et al., 2019)	44.79
BERT + GENIA (Chapter 7)	45.19
BioBERT (Lee et al., 2020)	44.33
Hard-matching	35.29
Soft-matching	41.95

Table 8.2: Performances on BioDRB across domains. Across domains means that the model is trained on PDTB and tested on BioDRB.

The results are shown in Table 8.2. The origin BERT achieves 44.79% accuracy across domains. The BERT model is fine tuned on PDTB and tested on BioDRB. With the same setting, the BERT + GENIA from Chapter 7 performs 45.19% and the gigantic in-domain pre-trained BioBERT achieves 44.33%. It is obvious to see that our proposed unsupervised method performs worse than the BERT-based methods on the whole. However, even though that the performances of the related methods we are comparing here are across domains, they are still trained on the task of implicit discourse relation classification on PDTB. This also indicates that BERT does not only care about the surface features of the tokens, it also has good ability in digging up the latent discourse relation patterns that are shared by the instances in different domains.

Nonetheless, our proposed unsupervised method achieves 35.29% with hard-matching and 41.95% with soft-matching, which is also comparable and competitive. This is because that the model didn't have access to supervised information other than the raw texts. Our assumption here is that statistically if both instances are talking about the similar entities (subjects and objects), it would be highly possible that they have the same discourse relation. In particularly, comparing the hard and soft matching variants, we believe that with a relatively loose constraint on identifying similar entities, the proposed method has better robustness and more reliable majority vote.

8.2.3 Conclusion and discussion

In this section, we introduce an unsupervised method with the help of the external entity information from the information retrieval system. Comparing with the supervised Bert-based models on the across-domains setting, although our experimental results haven't outperformed them, it achieved around 42% accuracy on BioDRB, only 2.33% worse than the gigantic in-domain pre-trained language model BioBert. We believe that it is due to the fact that the proposed method hasn't got any supervised information regarding the deep semantic representation and the discourse relation sense tags. It also motivates us to inject the extracted external entity information into the pre-trained model for supervised learning in the next section.

8.3 With pre-trained entity-augmented models

In Chapter 7, we have shown that the BERT is a crucial component in the high performance models. But it still needs more domain-specific knowledge to better encode the sentences from a new domain. With the pipeline proposed in the above section, now for each of the instance in the BioDRB, we have several related SPO triples from the explicit examples that are chosen with soft-matching. We here employ the recent proposed **Knowledge-enabled Language Representation model** (Liu et al., 2020, K-BERT) to integrate the external entity knowledge into the pre-trained language model for better argument representations.

8.3.1 K-BERT

Due to the domain discrepancies between the pre-training and fine-tuning, the unsupervised language models such as Bert etc. do not perform well on knowledge-driven tasks. Our previous results have also verified the conclusion. Integrating domain specific knowledge into pre-trained model can alleviate this problem. However, the process of knowledge acquisition can be inefficient and expensive (Liu et al., 2020).

In order to tackle the heterogeneous embedding space and knowledge noise problems, Liu et al. (2020) proposed a Knowledge-enabled Bidirectional Encoder Representation from

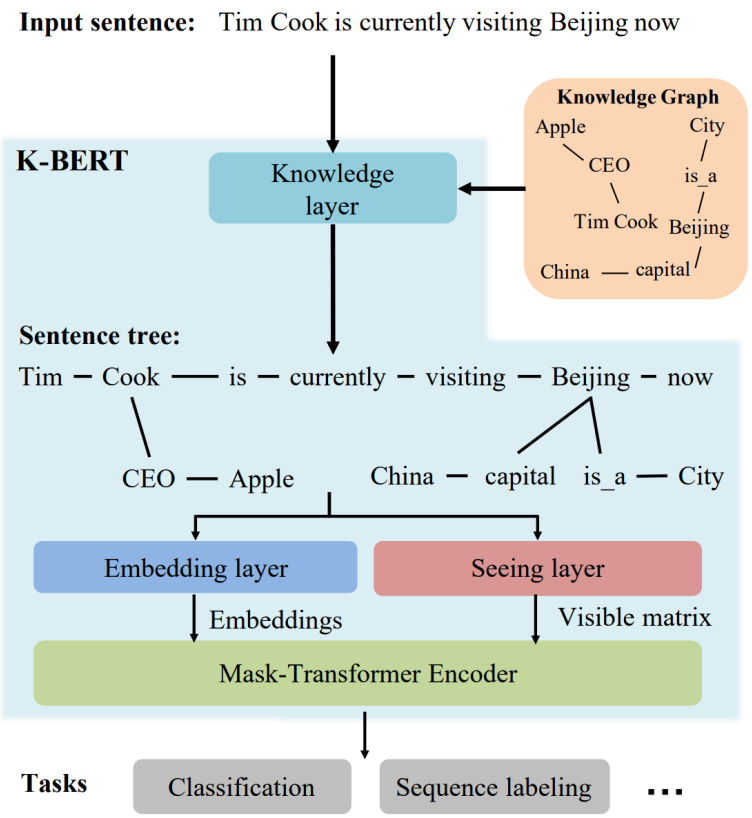

Figure 8.2: The structure of K-BERT. It is equipped with an editable knowledge graph which can be adapted to its application domain. Picture taken from Liu et al. (2020).

Transformers (K-BERT), as illustrated in Figure 8.2. It consists of a knowledge layer, embedding layer, seeing layer and mask-transformer. With the knowledge layer and the external knowledge graph, the input sentence has been expanded into a sentence tree and been input into the embedding layer and seeing layer. A seeing layer is to generate the visible matrix which controls the visible areas of each token to prevent changing the proper sequential order of the original sentence.

Figure 8.3 has a detailed demonstration about converting a sentence tree into the embedding representations. The whole sentence tree has been flattened into a sequence with the position index. The visible matrix is generated to keep the interactions of each of the tokens within the original sentence and also inside the knowledge graph triples. For example, with the hard-position index (the grey index on Figure 8.3), when predicting the token "Apple" which is an entity from the external KG, the visible matrix controls the self-attention layers in the transformer not to look into tokens other than "Tim Cook CEO Apple".

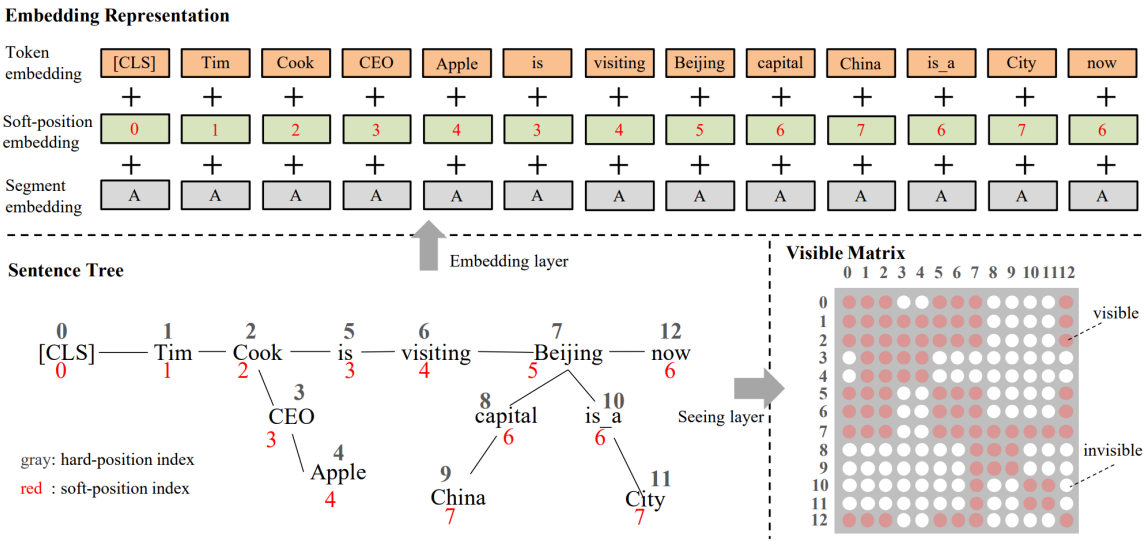

Figure 8.3: The process of converting a sentence tree into an embedding representation and a visible matrix. Picture taken from Liu et al. (2020).

8.3.2 Experiments and results

In Section 8.2, we introduce a pipeline to extract SPO triples from the explicit instances that are closely connected to the query instance with the proposed soft matching strategy. We see those SPO triples very good replacements for the knowledge graph which is used in the K-BERT. Comparing with the general knowledge graph, our extracted SPO triples have attached more importance on the discourse relations and stay in the same topic as they are extracted from the explicit discourse relation instances and have been pair matched with the entities in the input sentence / query.

For K-BERT pre-training, it is configured to the same parameter settings as the origin Google BERT (Devlin et al., 2019). And no additional knowledge graph are added to K-BERT during the pre-training phase, because it would make the word vectors of the linked entities too close or even the same to each other (Liu et al., 2020).

Therefore, for each input sentence, we attach the top 2 (default number from the K-BERT) similar SPO triples to the entities and convert it into a sentence tree, same as the work flow in Figure 8.2. We run the K-BERT as a classification task and evaluate it with the accuracy as per the conventional settings in previous chapter. The experimental results are illustrated in Table 8.3. We compare the results with some strong baselines that are reported in the recent

Methods	In-domain
Bai and Zhao (2018)	55.90
Bi-LSTM + Word2Vec	46.49
BERT (Devlin et al., 2019)	63.02
BERT + GENIA (Chapter 7)	66.04
BioBERT (Lee et al., 2020)	67.58
K-Bert (Liu et al., 2020)	**69.57**

Table 8.3: Performances on BioDRB within domain. Within domain here means 5-folds cross validation on BioDRB.

work. K-BERT, which is initialized with the original BERT parameters, achieves 69.57% accuracy on the 5 folds cross validation on BioDRB and outperforms the gigantic in-domain continuously pre-trained BioBERT with around 2 percents.

In addition, because of the extra external SPO triples, the K-BERT outperforms the original pure BERT by 6.5%. It also verifies the benefits brought by having the relevant SPO entities within the sentence when encoding.

8.3.3 Conclusion and discussion

In this section, we use the SPO triples extracted from Section 8.2 and employ the entity-augmented pre-trained language model K-BERT for the implicit discourse relation classification on BioDRB. The experimental results show that with the relevant entities, the representation from the model can expand the arguments with more entity-enhanced information and attach more importance to the interaction between the entities within the two relational arguments. The entity-enhanced language model achieves the state of the art results on BioDRB, outperforms the origin BERT and continuously pre-trained BioBERT with a significant margin.

8.4 Summary

In this chapter, we aim to integrate the entity information into the pre-trained language models to augment the model with more entity-specific information in the argument's encoding. We firstly propose the unsupervised majority voting system, which is motivated on

the assumption that if both discourse instances (both implicit or explicit ones) are talking about the similar entities, there is high possibility that the discourse relations entailed in them are the same. The results demonstrate competitive results on the BioDRB without any label supervision in the whole pipeline.

However, the proposed unsupervised method has failed comparing with the supervised methods across domains. This means that even in the very different domain, training on the implicit discourse relation task still helps the model to grasp some latent patterns or differences between the different discourse relations. We also show that with the extra information brought by the extracted SPO triples, the entity-enhanced pre-trained language model K-BERT shows more advantages in dealing with the professional and rare words in the new Biomedical domain. It achieves the state of the art performance on BioDRB, compared with the original BERT and the gigantic continuously in-domain pre-trained BioBERT with 6.5% and 2.0% accuracy respectively.

Even though that there are plenty of things, such as joint training with multi-tasks and combining cognitive models into the pre-trained language models, that can be done with the supervised learning methods, especially with the knowledge augmented methods like K-BERT (Liu et al., 2020), KG-BERT (Yao et al., 2019) and KnowBERT (Peters et al., 2019) etc. We would like to leave that to the future work. In the next chapter, we will conclude the whole dissertation and have an outlook of promising research directions to the task of implicit discourse relation classification.

Chapter 9

Conclusion and Outlook

In this thesis, we mainly focus on how to improve the performance of implicit discourse relation classification, which is to recognize the implicit discourse relation given the two arguments. Due to the lack of informative cues like the explicit connectives between the arguments, the task has been the bottleneck of the discourse parser. We tackle this task from different angles that are motivated by the weaknesses of previous proposed methods, and propose new approaches respectively. We evaluate the proposed methods on the manually annotated corpora such as the Penn Discourse Treebank (PDTB) and Biomedical Discourse Relation Bank (BioDRB), the results verify the effectiveness on different data-split settings.

In this chapter, we recap the contributions of our research work in this dissertation. We briefly discuss the problems we met in improving the classifier's performance and the solutions we propose. At last, we will conclude this dissertation with some directions for future work.

9.1 Conclusion

This thesis makes four main contributions to the research in implicit discourse relation classification. They include:

1. Illustrating the limited data problem and the risks in concluding upon them. (Chapter 3)

2. Acquiring automatically annotated data via the explicitation process during human translation between English and other languages. (Chapter 4, 5)

3. Better representation of discourse relation arguments. (Chapter 6, 7)

4. Entity enhancement with an unsupervised method and pre-trained language model. (Chapter 8)

Limited Data Problem. Previous work on implicit discourse relation classification tend to use the conventional settings on the data split of the PDTB, such as using the section 2-21, 22, 23 as the training, validation and testing sets respectively, following the setting in Lin et al. (2009). This results in having only less than 800 implicit discourse instances in the test set. In order to figure out the risk in concluding upon the small size of test set, we employ a simple neural network model to concatenate the representation and the surface features to predict the implicit discourse relation on both the conventional setting and the cross validation. The experiment results suggest that it comes to very different conclusions if actually running cross validation on all the sections, which means that the standard test section of the PDTB is way too small to draw conclusions about whether a feature is generally beneficial to this task or not, especially when using a relatively large label sets (11-ways classification here). This also motivates us to propose new approaches to acquire more annotated data for training in the later work.

Acquiring Annotated Data via the Explicitation between Language Pairs. Parallel corpora have been broadly used in training neural machine translation systems in the recent years. They are easy to access and are manually translated by human translators. Inspired by the fact that human translators often insert connectives to remove the ambiguity in the target language, we automatically back-translate the other languages (German, French, Czech) back to English. With the trained PDTB parser, it is easy to identify the explicit instances in the back-translations and label the original implicit English ones with the explicit label. In this work, we firstly try to use English-French pairs that are sentence-aligned. (Chapter 4) However, the problem of lacking topic consistency in the sentence-aligned corpora makes

it hard to extract the inter-sentential instances. In Chapter 5, we expand the idea to three language pairs. Both the experiments of the two chapters show that with the additional automatically labeled data for training, the performances on the test set have been improved, compared with the baseline models. In addition, we also find some interesting cases, patterns and cues in the human translation, especially with respect to the discourse relations and connectives.

Better Representation of Discourse Relation Arguments. When annotating the PDTB, the annotators are firstly asked to insert a connective and then label the instance with discourse relations. This is proven to be more efficient and accurate for the manual annotation. To better use the connective annotations and also motivated by the methodology of explicitation across languages proposed above, we use a sequence to sequence model to mimic the process of explicitation. It consists of three components: Encoder, Decoder and the Memory network. In addition to the typical relation classification task, we also train the model to predict the implicit connectives along with the arguments. With both the representations from the encoder and decoder, a gate serves as a selector to decide how much information from both sides to form the final representation of the implicit discourse instance. After combining the context vector from the memory network with the final representation, the model is able to predict the discourse relations. We evaluate the proposed model on the first level 4-ways, one-versus-all binary classification tasks and also the more fine-grained second level 11-ways classification, and the results show that with the seq2seq network, the model learns to grasp the key information and try to focus on the important words to help with the task when predicting the connectives. The model achieved the state of the art performances on the different settings at the time of the research.

However, a good implicit discourse classifier should not only be able to have good encoding ability, but also encode discourse expectation and learn typical temporal sequences, causes, consequences for all kinds of events. In Chapter 7, we find that with the help of the subtask "next sentence prediction", BERT has very powerful ability within and across domains. The results also show that removing the subtask hurts the performances of implicit discourse relation classification on different settings.

Entity enhancement with an unsupervised method and pre-trained language model. In

the Chapter 7, BERT shows very impressive ability in encoding both the sentences and discourse expectation. However, it also shows limited ability in keeping improving the performance with more in-domain raw texts for pre-training. This motivates us to integrate more domain-specific knowledge into the pre-trained model. We propose an unsupervised method with a information retrieval system and a quasi knowledge graphs system. With the proposed pipeline, we can extract large amount of explicit relation instances that are closely connected to the query (the implicit discourse relation instance to be labeled). After voting by those explicit relations, the query is labelled unsupervisedly and the accuracy is quite competitive comparing with the baseline BERT models. Despite that the BERT based methods have shown impressive abilities in achieving high performances, when it comes with a new domain, the ability is still limited. With all the SPO triples we extract with the proposed pipeline, we also integrate them into the pre-trained language model with the recent proposed K-BERT, achieve 69.57% accuracy and outperform the gigantic pre-trained BioBERT with 2% on BioDRB.

9.2 Outlook

The task of implicit discourse relation classification has gained increasing attention from the community recently. More and more researchers make lots of attempts to solve the problems in different aspects. However, there is still plenty room for significant improvements on this task as the overall performance is still low. Here we propose some future research directions based on the work we present and what we have learned through the path of thesis.

9.2.1 Multi-task Learning

Multi-task learning is an effective approach to improve the performance of a single task with the help of other related tasks. Mixed objectives can improve on the generalization ability of neural networks and result in better sentence representation and performance on one or more of the tasks. In Chapter 6, we use the argument generation as a secondary task to have better ability in sentence encoding and classification performance as well. Liu et al. (2016) combine the objectives of predicting connectives annotated for implicit relations, and

implicit vs. explicit relation label predictions. The core idea in these approaches is to improve learning by accessing other tasks for which larger amount of training data are available, and which require encoding the discourse relation arguments in ways that are also relevant for the prime task.

Sentiment polarity is one of the important features in training the statistical machine learning systems for implicit discourse relation recognition (Pitler et al., 2009). We believe that sentiment analysis is one of the relevant tasks to the discourse relation prediction. For instance, knowing the sentiment polarities of the two arguments is helpful for identifying contrastive relations. In addition, the task of question-answering can also be used in this setting. QA tasks can be related to discourse relation classification (Verberne et al., 2007; Chai and Jin, 2004), especially with respect to why-questions and causality in discourse relations.

Therefore, we believe the multi-task learning for implicit discourse relation classification is a research direction that deserves investigation in the future.

9.2.2 Connective Generation

When annotating the current largest available corpus PDTB, the annotators were asked to insert a proper connective before assigning a discourse relation to the arguments. This not only can make the annotators to have better understanding of the instance, but also is beneficial for higher inter-annotator agreement. Previous studies have tackled the task of predicting whether a discourse relation should be marked by an explicit discourse cue or whether the relation should instead be implicit (Yung et al., 2017; Patterson and Kehler, 2013), and have shown that information-theoretic considerations are relevant to this choice. Recently the pre-trained language models, such as GPT-2 (Radford et al., 2019), GPT-3 (Brown et al., 2020) etc., have shown very good abilities in generating coherent natural language sentences. However, there is no system to date that is able to predict which discourse connective among the ones that express the target discourse relation should be chosen. We believe that with a correctly predicted connective, it would be much easier for the model to predict the implicit discourse relations and can be later used as the realizer component of a natural language generation system to have more coherent generations.

9.2.3 Distant Supervision

Supervised learning methods require golden annotations for the model's training. The scale of annotated data becomes larger and larger due to the increasing number of parameters to be trained in the neural network models. Those annotations are mainly obtained through expert annotation, crowd-sourcing, or automatically labeling. It is difficult, expensive and time-consuming to get reliable golden annotations, especially for tasks like implicit discourse relations. For example, some world knowledge or professional background are required when trying to detect the causal relations between two sentences. Manually annotation can only alleviate the current situation when dealing with one single task, and when a more generalized method is needed, the annotation would be hard and even unfeasible to get. So we believe that the future efforts should focus on using more general data for distant supervision.

One way is similar to our work in Chapter 4 and 5. Our proposed pipeline have shown the effectiveness of using raw and rich resources to get automatically labelled data. However, it also needs better machine translation system to translate the connectives, better discourse parser to identify the spans and connectives of the explicit instances. With the rich and easy-to-access data and better strategy/criteria, we believe it would be promising to harvest more data for the implicit discourse relation classification task.

Another way is spiritually similar to the idea of Generative Adversarial Network (GAN) from the area of computer vision. The model is trained to distinguish golden label from noisy ones while generating noisy labels at the meantime. In this way, the model would be generalized to have better robustness in identifying the implicit labels. We believe this is a research direction that deserves more investigation in the future.

List of Figures

List of Tables

Bibliography

Amal Al-Saif and Katja Markert. The leeds Arabic discourse treebank: Annotating discourse connectives for Arabic. In *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10)*, Valletta, Malta, May 2010. European Language Resources Association (ELRA).

Gabor Angeli, Melvin Jose Johnson Premkumar, and Christopher D Manning. Leveraging linguistic structure for open domain information extraction. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 344–354, 2015.

Nicholas Asher, Nicholas Michael Asher, and Alex Lascarides. *Logics of conversation*. Cambridge University Press, 2003.

Fatemeh Torabi Asr. An information theoretic approach to production and comprehension of discourse markers. *PhD Dissertation*, 2015. doi: http://dx.doi.org/10.22028/D291-26632.

Fatemeh Torabi Asr and Vera Demberg. Implicitness of discourse relations. In *Proceedings of COLING 2012*, pages 2669–2684, Mumbai, India, 2012. The COLING 2012 Organizing Committee.

Fatemeh Torabi Asr and Vera Demberg. Uniform surprisal at the level of discourse relations: Negation markers and discourse connective omission. In *Proceedings of the 11th International Conference on Computational Semantics*, pages 118–128, 2015.

Hongxiao Bai and Hai Zhao. Deep enhanced representation for implicit discourse relation recognition. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 571–583, 2018.

Samy Bengio, Oriol Vinyals, Navdeep Jaitly, and Noam Shazeer. Scheduled sampling for sequence prediction with recurrent neural networks. In *Proceeding of NIPS*, pages 1171–1179, 2015.

Yoshua Bengio and Yves Grandvalet. Bias in estimating the variance of k-fold cross-validation. In *Statistical modeling and analysis for complex data problems*, pages 75–95. Springer, 2005.

Yoshua Bengio, Patrice Simard, and Paolo Frasconi. Learning long-term dependencies with gradient descent is difficult. *IEEE transactions on neural networks*, 5(2):157–166, 1994.

Yoshua Bengio, Réjean Ducharme, Pascal Vincent, and Christian Jauvin. A neural probabilistic language model. *Journal of machine learning research*, 3(Feb):1137–1155, 2003.

Taylor Berg-Kirkpatrick, David Burkett, and Dan Klein. An empirical investigation of statistical significance in nlp. In *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, pages 995–1005, Jeju Island, Korea, 2012. Association for Computational Linguistics.

Or Biran and Kathleen McKeown. Aggregated word pair features for implicit discourse relation disambiguation. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics*, pages 69–73, Sofia, Bulgaria, 2013. Association for Computational Linguistics.

John Blitzer, Ryan McDonald, and Fernando Pereira. Domain adaptation with structural correspondence learning. In *Proceedings of the 2006 conference on empirical methods in natural language processing*, pages 120–128, 2006.

Danushka Bollegala, Takanori Maehara, and Ken-ichi Kawarabayashi. Unsupervised cross-domain word representation learning. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 730–740, 2015.

Chloé Braud and Pascal Denis. Comparing word representations for implicit discourse relation classification. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 2201–2211, 2015.

Peter F Brown, Peter V Desouza, Robert L Mercer, Vincent J Della Pietra, and Jenifer C Lai. Class-based n-gram models of natural language. *Computational linguistics*, 18(4): 467–479, 1992.

Tom B Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *arXiv preprint arXiv:2005.14165*, 2020.

Deng Cai and Hai Zhao. Pair-aware neural sentence modeling for implicit discourse relation classification. In *International Conference on Industrial, Engineering and Other Applications of Applied Intelligent Systems*, pages 458–466. Springer, 2017.

Lynn Carlson and Daniel Marcu. Discourse tagging reference manual. *ISI Technical Report ISI-TR-545*, 54:56, 2001.

Lynn Carlson, Daniel Marcu, and Mary Ellen Okurovsky. Building a discourse-tagged corpus in the framework of rhetorical structure theory. In *Proceedings of the Second SIGdial Workshop on Discourse and Dialogue*, 2001. URL `https://www.aclweb.org/anthology/W01-1605`.

Bruno Cartoni, Sandrine Zufferey, Thomas Meyer, and Andrei Popescu-Belis. How comparable are parallel corpora? measuring the distribution of general vocabulary and connectives. In *Proceedings of the 4th Workshop on Building and Using Comparable Corpora: Comparable Corpora and the Web*, pages 78–86. Association for Computational Linguistics, 2011.

Bruno Cartoni, Sandrine Zufferey, and Thomas Meyer. Annotating the meaning of discourse connectives by looking at their translation: The translation-spotting technique. *D&D*, 4 (2):65–86, 2013.

Joyce Chai and Rong Jin. Discourse structure for context question answering. In *Proceedings of the Workshop on Pragmatics of Question Answering at HLT-NAACL 2004*, pages 23–30, 2004.

Jifan Chen, Qi Zhang, Pengfei Liu, Xipeng Qiu, and Xuanjing Huang. Implicit discourse relation detection via a deep architecture with gated relevance network. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 1726–1735, Berlin, Germany, 2016. Association for Computational Linguistics.

Kyunghyun Cho, Bart Van Merriënboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. Learning phrase representations using rnn encoder-decoder for statistical machine translation. *arXiv preprint arXiv:1406.1078*, 2014.

Junyoung Chung, Caglar Gulcehre, KyungHyun Cho, and Yoshua Bengio. Empirical evaluation of gated recurrent neural networks on sequence modeling. *arXiv preprint arXiv:1412.3555*, 2014.

Philipp Cimiano, Uwe Reyle, and Jasmin Šarić. Ontology-driven discourse analysis for information extraction. *Data & Knowledge Engineering*, 55(1):59–83, 2005.

Luciano Del Corro and Rainer Gemulla. Clausie: clause-based open information extraction. In *Proceedings of the 22nd international conference on World Wide Web*, pages 355–366, 2013.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, 2019.

Bhuwan Dhingra, Zhilin Yang, William W Cohen, and Ruslan Salakhutdinov. Linguistic knowledge as memory for recurrent neural networks. *arXiv preprint arXiv:1703.02620*, 2017.

Shima Gerani, Yashar Mehdad, Giuseppe Carenini, Raymond T. Ng, and Bita Nejat. Abstractive summarization of product reviews using discourse structure. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing*, pages 1602–1613. Association for Computational Linguistics, 2014.

Barbara J Grosz and Candace L Sidner. Attention, intentions, and the structure of discourse. *Computational linguistics*, 12(3):175–204, 1986.

Honglei Guo, Huijia Zhu, Zhili Guo, Xiaoxun Zhang, Xian Wu, and Zhong Su. Domain adaptation with latent semantic association for named entity recognition. In *Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pages 281–289, 2009.

Francisco Guzmán, Shafiq R. Joty, Lluís Màrquez, and Preslav Nakov. Using discourse structure improves machine translation evaluation. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics*, pages 687–698. Association for Computational Linguistics, 2014.

Ruifang He, Jian Wang, Fengyu Guo, and Yugui Han. Transs-driven joint learning architecture for implicit discourse relation recognition. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 139–148, 2020.

Hugo Hernault, Danushka Bollegala, and Mitsuru Ishizuka. A semi-supervised approach to improve classification of infrequent discourse relations using feature vector extension. In *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*, pages 399–409. Association for Computational Linguistics, 2010.

Christopher Hidey and Kathleen McKeown. Identifying causal relation using parallel wikipedia articles. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics*, pages 1424–1433. Association for Computational Linguistics, 2016.

Jerry R Hobbs. Coherence and coreference. *Cognitive science*, 3(1):67–90, 1979.

Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural computation*, 9(8):1735–1780, 1997.

Jet Hoek and Sandrine Zufferey. Factors influencing the implicitation of discourse relations across languages. In *Proceedings 11th Joint ACL-ISO Workshop on Interoperable Semantic Annotation (isa-11)*, pages 39–45. TiCC, Tilburg center for Cognition and Communication, 2015.

Peter Jansen, Mihai Surdeanu, and Peter Clark. Discourse complements lexical semantics for non-factoid answer reranking. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, volume 1, pages 977–986, 2014.

Yangfeng Ji and Jacob Eisenstein. Representation learning for text-level discourse parsing. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 13–24, 2014.

Yangfeng Ji and Jacob Eisenstein. One vector is not enough: Entity-augmented distributed semantics for discourse relations. *Transactions of the Association of Computational Linguistics*, 3(1):329–344, 2015.

Yangfeng Ji, Gholamreza Haffari, and Jacob Eisenstein. A latent variable recurrent neural network for discourse relation language models. In *Proceedings of NAACL-HLT*, pages 332–342, 2016.

Amina Kadry and Laura Dietz. Open relation extraction for support passage retrieval: Merit and open issues. In *Proceedings of the 40th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 1149–1152, 2017.

Nal Kalchbrenner, Edward Grefenstette, and Phil Blunsom. A convolutional neural network for modelling sentences. In *Proceedings of ACL*, 2014.

J-D Kim, Tomoko Ohta, Yuka Tateisi, and Jun'ichi Tsujii. Genia corpus—a semantically annotated corpus for bio-textmining. *Bioinformatics*, 19(suppl_1):i180–i182, 2003.

Ji-Hyun Kim. Estimating classification error rate: Repeated cross-validation, repeated hold-out and bootstrap. *Computational Statistics & Data Analysis*, volume53(11):3735–3745, 2009.

Diederik Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

Yudai Kishimoto, Yugo Murawaki, and Sadao Kurohashi. A knowledge-augmented neural network model for implicit discourse relation classification. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 584–595, 2018.

Guillaume Klein, Yoon Kim, Yuntian Deng, Jean Senellart, and Alexander M. Rush. Opennmt: Open-source toolkit for neural machine translation. *arXiv preprint arXiv:1701.02810*, 2017.

Philipp Koehn. Europarl: A parallel corpus for statistical machine translation. In *MT summit*, volume 5, pages 79–86, 2005.

Philipp Koehn, Hieu Hoang, Alexandra Birch, Chris Callison-Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowan, Wade Shen, Christine Moran, Richard Zens, et al. Moses: Open source toolkit for statistical machine translation. In *Proceedings of the 45th annual meeting of the ACL on interactive poster and demonstration sessions*, pages 177–180. Association for Computational Linguistics, 2007.

Moshe Koppel and Noam Ordan. Translationese and its dialects. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics*, pages 1318–1326. Association for Computational Linguistics, 2011.

Kerstin Kunz and Ekaterina Lapshinova-Koltunski. Cross-linguistic analysis of discourse variation across registers. *Nordic Journal of English Studies*, 14(1):258–288, 2015.

Majid Laali and Leila Kosseim. Inducing discourse connectives from parallel texts. In *Proceedings of the 25th International Conference on Computational Linguistics (COLING-2014)*, pages 610–619, 2014.

Man Lan, Yu Xu, Zheng-Yu Niu, et al. Leveraging synthetic discourse data via multi-task learning for implicit discourse relation recognition. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics*, pages 476–485. Association for Computational Linguistics, 2013.

Man Lan, Jianxiang Wang, Yuanbin Wu, Zheng-Yu Niu, and Haifeng Wang. Multi-task attention-based neural networks for implicit discourse relationship representation and identification. In *Proceedings of EMNLP*, pages 1299–1308, 2017.

Jinhyuk Lee, Wonjin Yoon, Sungdong Kim, Donghyeon Kim, Sunkyu Kim, Chan Ho So, and Jaewoo Kang. Biobert: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*, 36(4):1234–1240, 2020.

Beth Levin. *English verb classes and alternations: A preliminary investigation*. University of Chicago press, 1993.

Qi Li. Literature survey: domain adaptation algorithms for natural language processing. *Department of Computer Science The Graduate Center, The City University of New York*, pages 8–10, 2012.

Maria Liakata, Simon Dobnik, Shyamasree Saha, Colin Batchelor, and Dietrich Rebholz-Schuhmann. A discourse-driven content model for summarising scientific articles evaluated in a complex question answering task. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 747–757, Seattle, Washington, USA, October 2013. Association for Computational Linguistics. URL `https://www.aclweb.org/anthology/D13-1070`.

Ziheng Lin, Min-Yen Kan, and Hwee Tou Ng. Recognizing implicit discourse relations in the penn discourse treebank. In *Proceedings of EMNLP*, pages 343–351, 2009.

Ziheng Lin, Hwee Tou Ng, and Min-Yen Kan. A pdtb-styled end-to-end discourse parser. *Natural Language Engineering*, 20(02):151–184, 2014.

Qi Liu, Yue Zhang, and Jiangming Liu. Learning domain representation for multi-domain sentiment classification. In *Proceedings of NAACL*, pages 541–550, 2018.

Weijie Liu, Peng Zhou, Zhe Zhao, Zhiruo Wang, Qi Ju, Haotang Deng, and Ping Wang. K-bert: Enabling language representation with knowledge graph. In *AAAI*, pages 2901–2908, 2020.

Yang Liu and Sujian Li. Recognizing implicit discourse relations via repeated reading: Neural networks with multi-level attention. In *Proceedings of EMNLP*, pages 1224–1233, 2016.

Yang Liu, Sujian Li, Xiaodong Zhang, and Zhifang Sui. Implicit discourse relation classification via multi-task neural networks. In *Proceedings of AAAI*, pages 2750–2756, 2016.

Xiaolu Lu, Soumajit Pramanik, Rishiraj Saha Roy, Abdalghani Abujabal, Yafang Wang, and Gerhard Weikum. Answering complex questions by joining multi-document evidence with quasi knowledge graphs. In *Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 105–114, 2019.

Minh-Thang Luong, Hieu Pham, and Christopher D Manning. Effective approaches to attention-based neural machine translation. In *Proceedings of EMNLP*, pages 1412–1421, 2015.

William C Mann and Sandra A Thompson. Rhetorical structure theory: Toward a functional theory of text organization. *Text-interdisciplinary Journal for the Study of Discourse*, 8(3): 243–281, 1988.

Daniel Marcu and Abdessamad Echihabi. An unsupervised approach to recognizing discourse relations. In *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*, pages 368–375. Association for Computational Linguistics, 2002.

Mitchell Marcus, Grace Kim, Mary Ann Marcinkiewicz, Robert MacIntyre, Ann Bies, Mark Ferguson, Karen Katz, and Britta Schasberger. The penn treebank: annotating predicate argument structure. In *Proceedings of the workshop on Human Language Technology*, pages 114–119. Association for Computational Linguistics, 1994.

Mitchell P. Marcus, Beatrice Santorini, and Mary Ann Marcinkiewicz. Building a large annotated corpus of English: The Penn Treebank. *Computational Linguistics*, 19(2):313–330, 1993. URL `https://www.aclweb.org/anthology/J93-2004`.

Sameer Maskey and Julia Hirschberg. Comparing lexical, acoustic/prosodic, structural and discourse features for speech summarization. In *Ninth European Conference on Speech Communication and Technology*, 2005.

Mausam Mausam. Open information extraction systems and downstream applications. In *Proceedings of the twenty-fifth international joint conference on artificial intelligence*, pages 4074–4077, 2016.

Thomas Meyer, Najeh Hajlaoui, and Andrei Popescu-Belis. Disambiguating discourse connectives for statistical machine translation. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 23(7):1184–1197, 2015.

Claudiu Mihăilă and Sophia Ananiadou. Semi-supervised learning of causal relations in biomedical scientific discourse. *Biomedical engineering online*, 13(2):1, 2014.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. Distributed representations of words and phrases and their compositionality. In *Proceedings of NIPS*, pages 3111–3119, 2013.

Allen Nie, Erin Bennett, and Noah Goodman. Dissent: Learning sentence representations from explicit discourse relations. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4497–4510, 2019.

Joonsuk Park and Claire Cardie. Improving implicit discourse relation recognition through feature set optimization. In *13th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, page 108, 2012.

Gary Patterson and Andrew Kehler. Predicting the presence of discourse connectives. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 914–923, 2013.

Jeffrey Pennington, Richard Socher, and Christopher Manning. Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 1532–1543, 2014.

Matthew Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. Deep contextualized word representations. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 2227–2237, 2018.

Matthew E. Peters, Mark Neumann, Robert L Logan, Roy Schwartz, Vidur Joshi, Sameer Singh, and Noah A. Smith. Knowledge enhanced contextual word representations. In *EMNLP*, 2019.

Emily Pitler, Mridhula Raghupathy, Hena Mehta, Ani Nenkova, Alan Lee, and Aravind Joshi. Easily identifiable discourse relations. In *Coling 2008: Companion volume: Posters*, pages 87–90, 2008.

Emily Pitler, Annie Louis, and Ani Nenkova. Automatic sense prediction for implicit discourse relations in text. In *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP: Volume 2-Volume 2*, pages 683–691, Suntec, Singapore, 2009. Association for Computational Linguistics.

Rashmi Prasad, Nikhil Dinesh, Alan Lee, Eleni Miltsakaki, Livio Robaldo, Aravind Joshi, and Bonnie Webber. The penn discourse treebank 2.0. In *Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC08)*, 2008.

Rashmi Prasad, Susan McRoy, Nadya Frid, Aravind Joshi, and Hong Yu. The biomedical discourse relation bank. *BMC bioinformatics*, 12(1):188, 2011.

Lianhui Qin, Zhisong Zhang, and Hai Zhao. Implicit discourse relation recognition with context-aware character-enhanced embeddings. In *Proceedings of COLING*, 2016a.

Lianhui Qin, Zhisong Zhang, and Hai Zhao. A stacking gated neural architecture for implicit discourse relation classification. In *Proceedings of EMNLP*, pages 2263–2270, 2016b.

Lianhui Qin, Zhisong Zhang, Hai Zhao, Zhiting Hu, and Eric Xing. Adversarial connective-exploiting networks for implicit discourse relation classification. In *Proceedings of the 55th*

Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), volume 1, pages 1006–1017, 2017.

Ella Rabinovich, Shuly Wintner, and Ofek Luis Lewinsohn. The haifa corpus of translationese. *arXiv preprint arXiv:1509.03611*, 2015.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. Language models are unsupervised multitask learners. *OpenAI Blog*, 1(8):9, 2019.

Ines Rehbein, Merel Scholman, and Vera Demberg. Annotating discourse relations in spoken language: A comparison of the pdtb and ccr frameworks. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)*, pages 23–28, Portorož, Slovenia, 2016. European Language Resources Association.

Hannah Rohde, Anna Dickinson, Chris Clark, Annie Louis, and Bonnie Webber. Recovering discourse relations: Varying influence of discourse adverbials. In *Proceedings of the First Workshop on Linking Computational Models of Lexical, Sentential and Discourse-level Semantics*, pages 22–31, 2015.

Samuel Rönnqvist, Niko Schenk, and Christian Chiarcos. A recurrent neural model with attention for the recognition of chinese implicit discourse relations. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 256–262, Vancouver, Canada, July 2017. Association for Computational Linguistics.

Attapol Rutherford and Nianwen Xue. Discovering implicit discourse relations through brown cluster pair representation and coreference patterns. In *Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics*, pages 645–654. Association for Computational Linguistics, 2014.

Attapol Rutherford and Nianwen Xue. Improving the inference of implicit discourse relations via classifying explicit discourse connectives. In *Proceedings of NAACL*, pages 799–808, 2015.

Attapol Rutherford, Vera Demberg, and Nianwen Xue. A systematic study of neural discourse models for implicit discourse relation. In *European Chapter of the Association*

for Computational Linguistics.(EACL), Valencia, Spain, 2017a. Association for Computational Linguistics.

Attapol Rutherford, Vera Demberg, and Nianwen Xue. A systematic study of neural discourse models for implicit discourse relation. In *Proceedings of EACL*, pages 281–291, 2017b.

Attapol Rutherford, Vera Demberg, and Nianwen Xue. A systematic study of neural discourse models for implicit discourse relation. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics*, pages 281–291. Association for Computational Linguistics, 2017c.

Wei Shi and Vera Demberg. On the need of cross validation for discourse relation classification. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*, pages 150–156, 2017.

Wei Shi and Vera Demberg. Learning to explicitate connectives with seq2seq network for implicit discourse relation classification. In *Proceedings of the 13th International Conference on Computational Semantics-Long Papers*, pages 188–199, 2019a.

Wei Shi and Vera Demberg. Next sentence prediction helps implicit discourse relation classification within and across domains. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 5794–5800, 2019b.

Wei Shi, Frances Yung, Raphael Rubino, and Vera Demberg. Using explicit discourse connectives in translation for implicit discourse relation classification. In *Proceedings of the Eighth International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 484–495, 2017.

Richard Socher, Alex Perelygin, Jean Wu, Jason Chuang, Christopher D Manning, Andrew Y Ng, and Christopher Potts. Recursive deep models for semantic compositionality over a sentiment treebank. In *Proceedings of the 2013 conference on empirical methods in natural language processing*, pages 1631–1642, 2013.

Radu Soricut and Daniel Marcu. Sentence level discourse parsing using syntactic and lexical information. In *Proceedings of the 2003 Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics*, pages 228–235, 2003.

Caroline Sporleder and Alex Lascarides. Using automatically labelled examples to classify rhetorical relations: An assessment. *Natural Language Engineering*, 14(3):369–416, 2008.

Philip J Stone, Dexter C Dunphy, and Marshall S Smith. *The General Inquirer: A Computer Approach to Content Analysis.* MIT press, 1966.

Kai Sheng Tai, Richard Socher, and Christopher D Manning. Improved semantic representations from tree-structured long short-term memory networks. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1556–1566, 2015.

Deborah Tannen, Heidi E Hamilton, and Deborah Schiffrin. *The handbook of discourse analysis*. John Wiley & Sons, 2015.

Sara Tonelli, Giuseppe Riccardi, Rashmi Prasad, and Aravind K Joshi. Annotation of discourse relations for conversational spoken dialogs. In *LREC*, 2010.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in neural information processing systems*, pages 5998–6008, 2017.

Suzan Verberne, Lou Boves, Nelleke Oostdijk, and Peter-Arno Coppen. Evaluating discourse-based answer extraction for why-question answering. In *Proceedings of the 30th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 735–736, 2007.

Yannick Versley. Discovery of ambiguous and unambiguous discourse connectives via annotation projection. In *AEPC*, pages 83–82, 2010.

Peilu Wang, Yao Qian, Frank K Soong, Lei He, and Hai Zhao. Word embedding for recurrent neural network based tts synthesis. In *Acoustics, Speech and Signal Processing (ICASSP), 2015 IEEE International Conference on*, pages 4879–4883. IEEE, 2015.

Xun Wang, Sujian Li, Jiwei Li, and Wenjie Li. Implicit discourse relation recognition by selecting typical training examples. In *Proceedings of the 24th International Conference on Computational Linguistics (COLING-2012)*, pages 2757–2772, 2012.

Changxing Wu, Xiaodong Shi, Yidong Chen, Yanzhou Huang, and Jinsong Su. Bilingually-constrained synthetic data for implicit discourse relation recognition. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 2306–2312. Association for Computational Linguistics, 2016.

Yu Xu, Man Lan, Yue Lu, Zheng Yu Niu, and Chew Lim Tan. Connective prediction using machine learning for implicit discourse relation classification. In *Neural Networks (IJCNN), The 2012 International Joint Conference on*, pages 1–8. IEEE, 2012.

Gui-Rong Xue, Wenyuan Dai, Qiang Yang, and Yong Yu. Topic-bridged plsa for cross-domain text classification. In *Proceedings of the 31st annual international ACM SIGIR conference on Research and development in information retrieval*, pages 627–634, 2008.

Nianwen Xue, Hwee Tou Ng, Sameer Pradhan, Rashmi Prasad, Christopher Bryant, and Attapol Rutherford. The conll-2015 shared task on shallow discourse parsing. In *Proceedings of the CoNLL-15 Shared Task*, pages 1–16. Association for Computational Linguistics, 2015.

Nianwen Xue, Hwee Tou Ng, Attapol Rutherford, Bonnie Webber, Chuan Wang, and Hongmin Wang. Conll 2016 shared task on multilingual shallow discourse parsing. In *Proceedings of the CoNLL-16 shared task*, pages 1–19. Association for Computational Linguistics, 2016.

Zhilin Yang, Zihang Dai, Yiming Yang, Jaime Carbonell, Russ R Salakhutdinov, and Quoc V Le. Xlnet: Generalized autoregressive pretraining for language understanding. In *Advances in neural information processing systems*, pages 5753–5763, 2019.

Liang Yao, Chengsheng Mao, and Yuan Luo. Kg-bert: Bert for knowledge graph completion. *arXiv preprint arXiv:1909.03193*, 2019.

Yasuhisa Yoshida, Jun Suzuki, Tsutomu Hirao, and Masaaki Nagata. Dependency-based discourse parser for single-document summarization. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1834–1839, 2014.

Frances Yung, Kevin Duh, Taku Komura, and Yuji Matsumoto. A psycholinguistic model for the marking of discourse relations. *Dialogue & Discourse*, 8(1):106–131, 2017.

Biao Zhang, Jinsong Su, Deyi Xiong, Yaojie Lu, Hong Duan, and Junfeng Yao. Shallow convolutional neural network for implicit discourse relation recognition. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 2230–2235, 2015.

Lanjun Zhou, Wei Gao, Bin Li, Zhong Wei, and Kam-Fai Wong. Cross-lingual identification of ambiguous discourse connectives for resource-poor language. In *Proceedings of COLING 2012: Posters*, pages 1409–1418. The COLING 2012 Organizing Committee, 2012.

Peng Zhou, Wei Shi, Jun Tian, Zhenyu Qi, Bingchen Li, Hongwei Hao, and Bo Xu. Attention-based bidirectional long short-term memory networks for relation classification. In *Proceedings of ACL*, pages 207–212, 2016.

Yuping Zhou and Nianwen Xue. The chinese discourse treebank: a chinese corpus annotated with discourse relations. *Language Resources and Evaluation*, 49(2):397–431, 2015.

Zhi-Min Zhou, Yu Xu, Zheng-Yu Niu, Man Lan, Jian Su, and Chew Lim Tan. Predicting discourse connectives for implicit discourse relation recognition. In *Proceedings of the 23rd International Conference on Computational Linguistics: Posters*, pages 1507–1514, Beijing, China, 2010. Association for Computational Linguistics.

Sandrine Zufferey. Discourse connectives across languages. *Languages in Contrast*, 16(2): 264–279, 2016.